Generating Mathematical Derivations with Large Language Models

Generating Mathematical Derivations with Large Language Models

Jordan Meadows[1], **Marco Valentino**[2], **André Freitas**[1,2]
[1]University of Manchester, United Kingdom
[2]Idiap Research Institute, Switzerland
jordan.meadows@postgrad.manchester.ac.uk
{marco.valentino, andre.freitas}@idiap.ch

Abstract

The derivation of mathematical results in specialised fields, using Large Language Models (LLMs), is an emerging research direction that can help identify models' limitations, and potentially support mathematical discovery. In this paper, we leverage a symbolic engine to generate derivations of equations at scale, and investigate the capabilities of LLMs when deriving goal equations from premises. Specifically, we employ in-context learning for GPT and fine-tune a range of T5 models to compare the robustness and generalisation of pre-training strategies to specialised models. Empirical results show that fine-tuned FLAN-T5-large (`MathT5`[1]) outperforms GPT models on all static and out-of-distribution test sets in conventional scores. However, an in-depth analysis reveals that the fine-tuned models are more sensitive to perturbations involving unseen symbols and (to a lesser extent) changes to equation structure. In addition, we analyse 1.7K equations, and over 200 derivations, to highlight common reasoning errors such as the inclusion of incorrect, irrelevant, and redundant equations. Finally, we explore the suitability of existing metrics for evaluating mathematical derivations and find evidence that, while they can capture general properties such as sensitivity to perturbations, they fail to highlight fine-grained reasoning errors and essential differences between models. Overall, this work demonstrates that training models on synthetic data may improve their math capabilities beyond much larger LLMs, but current metrics are not appropriately assessing the quality of generated mathematical text.

1 Introduction

Robust mathematical reasoning is a desirable emerging capability of Large Language Models (LLMs) (Wu et al., 2022). LLMs have been shown to resemble or surpass human performance in various settings, including undergraduate mathematics and physics question answering (Lewkowycz et al., 2022; Drori et al., 2022). However, recent assessments question their adequacy, particularly in sustained multi-hop reasoning (Frieder et al., 2023).

In this paper, we aim to provide an in-depth discussion centred around equation derivations, focusing on models' ability to learn and sequentially apply symbolic operations to premise equations, in order to derive goal equations defined within prompts. Such *derivation-style* equational reasoning (Plaisted, 1993; Premtoon et al., 2020) is at the core of many applied mathematical fields, such as theoretical physics and engineering. It is difficult to formalise (Kaliszyk et al., 2015), and is hence incompatible with theorem provers (Govindarajalulu et al., 2015; Hulette et al., 2015; Davis, 2019; Meadows and Freitas, 2021). Moreover, the granularity of the operations within a derivation is typically far greater than what is surfaced on published derivations (Mann et al., 2018), with many steps being omitted or summarised, which ultimately leads to a fundamental incompleteness problem for reasoning data available for training (Villalobos et al., 2022). Given that granular workings (often with implicit operations) contribute to much of the theoretical research distilled in papers, that generative models have been shown to hallucinate on domain-specific reasoning problems (Shuster et al., 2021; Taylor et al., 2022; Frieder et al., 2023; Wysocka et al., 2023), and that granular reasoning lends itself better to *explainability and inference control* (Hebenstreit et al., 2023; Yao et al., 2023; Yuan et al., 2023), it is clear we must extend the *fine-grained reasoning* abilities of language models. This is especially true if we aim to use them to reliably derive and explain results in specialised fields.

This work builds upon a recent framework for symbolic data generation and robustness evaluation (Meadows et al., 2023), that attempts to sys-

[1]https://huggingface.co/jmeadows17/MathT5-large

tematically emulate and perturb complex forms of equational reasoning. From this, we contribute:

(1.) An improved algorithm used to produce 15K procedurally generated LaTeX derivations and prompts, comprising between 4-10 equations, for fine-tuning and evaluating language models (and generative models in general) on fine-grained multi-step equational reasoning.

(2.) Out-of-disftribution test sets generated by applying systematic perturbations to a static test set comprising 2K examples.

(3.) We fine-tune T5 and FLAN-T5 models and employ ChatGPT and GPT-4 via in-context learning on the generated data, evaluating the models on static and perturbed sets, to assess their ability to generalise to mathematical derivations where targeted elements of reasoning have been altered.

(4.) We analyse derivations generated by fine-tuned FLAN-T5-large, ChatGPT, and GPT-4, and highlight reasoning failures of models in an extensive quantitative and qualitative analysis. The results show that while existing evaluation metrics score fine-tuned T5 models above GPT on many test sets, GPT is affected less by perturbations. In particular, the fine-tuned models struggle with out-of-distribution symbols and alternative equation structures.

Overall, this work suggests that fine-tuning smaller LLMs on synthetic derivations can yield detailed multi-step equational reasoning capabilities exceeding those of larger models (*e.g.,* GPT). Naturally, smaller models are less robust to out-of-distribution examples, but metrics do not reveal this from conventional evaluation procedures. We find that common text generation metrics feature a number of limitations in this regard, and we suggest further research developing more sophisticated metrics in this context. Related resources for reproducing and extending our experiments are available online[2].

2 Related Work

Our focus is the generation and perturbation of informal mathematical reasoning that resembles stepwise detailed human workings with LLMs (Brown et al., 2020; Ahmed and Devanbu, 2022; Song et al., 2022; Ge et al., 2023; Hu et al., 2023; Yang et al., 2023). While we presently consider *solely equations*, math generation exists in various forms, and is split between approaches that consider formal languages, and those that consider informal mathematical natural language (Meadows and Freitas, 2022; Ferreira and Freitas, 2020; Welleck et al., 2021; Ferreira et al., 2022; Valentino et al., 2022). In the formal case, GPT-j (Polu and Sutskever, 2020; Polu et al., 2022), LISA (Jiang et al., 2021), and Baldur (First et al., 2023) model Metamath and Isabelle/HOL proofs. For generation involving informal reasoning, an approach based on OpenAI's Codex (Chen et al., 2021; Drori et al., 2022) translates university-level problems into executable code, and generates solution explanations. Minerva (Lewkowycz et al., 2022) is a PaLM (Chowdhery et al., 2022) model trained on a large corpus of mathematical text, and solves university-level problems in applied math, outputting solutions in the form of mathematical natural language. NaturalProver (Welleck et al., 2022) generates similar solutions to proofs from a curated dataset (Welleck et al., 2021), and is most similar to our present work. However, we deviate in three ways. Firstly, we only generate chains of LaTeX equations and ignore natural language. Secondly, our prompts and derivations are procedurally generated, and are engineered for use with lightweight models, while containing up to 10 equations with complex symbols. Thirdly, our use of perturbations following a related approach (Meadows et al., 2023), allows for pairwise comparison between examples that differ by targeted aspects of equations, such as changes to structure and symbols. We further expand on this by improving the coherence of the reasoning generation algorithm, ensuring that derivations do not contain irrelevant steps, and consider more complex symbols (*e.g.,* g'_{ε}).

3 Derivation Generation with LLMs

Given a goal equation G and premise $\mathcal{P}$, that are arranged within some prompt template $t(\mathcal{P}, G)$, we aim to assess the ability of an LLM to systematically apply a set of symbolic operations $\mathcal{O}$ and generate a sequence of intermediate equations $\hat{\mathcal{D}}$ such that, starting from $\mathcal{P}$, $\hat{\mathcal{D}}$ represents a reasonable derivation of G. Given model $\mathcal{M}$, a derivation is generated through $\mathcal{M} : t(\mathcal{P}, G) \mapsto \hat{\mathcal{D}}$. Some idealised metric M^* scores a derivation through $M^* : (\mathcal{D}^*, \hat{\mathcal{D}}) \mapsto \mathcal{S}$, where $\mathcal{D}^*$ is some ideal derivation. Assuming a suitable prompt t, we generally aim to optimise

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}}; M^*\big((\mathcal{D}^*, \mathcal{M} : t(\mathcal{P}, G) \mapsto \hat{\mathcal{D}})\big).$$

[2] https://github.com/jmeadows17/deriving-equations-with-LLMs

However, we do not have access to ideal derivations $\mathcal{D}^*$ corresponding to templates $t(\mathcal{P}, G)$, nor ideal metric M^* suitable for scoring $\hat{\mathcal{D}}$. Instead, we employ a symbolic engine to *estimate* ground truth derivations to obtain $\tilde{\mathcal{D}}^*$ (Alg. 1). Moreover, we are evaluating over a sample of derivations. This means that, in practice, we are instead finding $\mathcal{M}^*$ such that

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}}; \frac{1}{N}\sum_{i=1}^{N} M\big((\tilde{\mathcal{D}}^*{}_i, \mathcal{M} : t(\mathcal{P}_i, G_i) \mapsto \hat{\mathcal{D}}_i)\big),$$

where N is the sample size. In this work, M is a canonical text generation metric (*e.g.,* ROUGE).

4 Data Generation

In Meadows et al. (2023), static derivations are generated and perturbed to form out-of-distribution test sets. These static and perturbed sets are used to determine the impact of perturbations on model performance, by comparing static scores with perturbed scores. In this paper, we propose an algorithm (Alg. 1) to generate complete synthetic derivations that are used to build datasets for systematically assessing the performance of LLMs on the derivation generation task, including prompts and references.

4.1 Synthetic Derivation Generation

$$\mathrm{v_t}\left(\sigma_x, \hat{p}_0\right) = e^{\hat{p}_0 - \sigma_x}$$

$$-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right) = -\sigma_x + e^{\hat{p}_0 - \sigma_x}$$

$$\int \left(-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)\, d\sigma_x = \int \left(-\sigma_x + e^{\hat{p}_0 - \sigma_x}\right)\, d\sigma_x$$

$$\int \left(-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)\, d\sigma_x = -\frac{\sigma_x^2}{2} + f - e^{\hat{p}_0 - \sigma_x}$$

$$\int \left(-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)\, d\sigma_x = -\frac{\sigma_x^2}{2} + f - \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)$$

$$\frac{\partial}{\partial \sigma_x}\int \left(-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)\, d\sigma_x = \frac{\partial}{\partial \sigma_x}\left(-\frac{\sigma_x^2}{2} + f - \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)$$

$$\frac{\partial^2}{\partial f \partial \sigma_x}\int \left(-\sigma_x + \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)\, d\sigma_x = \frac{\partial^2}{\partial f \partial \sigma_x}\left(-\frac{\sigma_x^2}{2} + f - \mathrm{v_t}\left(\sigma_x, \hat{p}_0\right)\right)$$

Figure 1: An example of a synthetic mathematical derivation used to assess derivation capabilities of LLMs. Three equations would be given in the corresponding prompt: the first equation is a *premise*, the fourth equation is an *intermediate step*, the last equation is the *goal equation*.

The high-level description of the reasoning generation involves a vocabulary of symbols and a set of explicit operations. A premise equation $\mathcal{P}$ is initially generated such as the first equation in Fig. 1. To build up derivations from $\mathcal{P}$, operations are randomly selected and symbolically applied (Meurer et al., 2017) to both sides of the premise equation, leading to new equations. During this process, smaller unconnected derivations may exist in the sequence that stem from existing equations. All suitable smaller derivations are integrated into a coherent sequence that represents the output derivation. Operands used for generating equations are randomly sampled from all possible symbols and sub-expressions defined in the full chain so far. The introduction of new symbols comes from the introduction of new premises or integration constants. Hyperparameters control the derivation length, a memory bias towards recent equations, and the frequency that certain operations are attempted (Appendix E). We improve a limitation where derivations may contain unused equations, and ensure a directed acyclic graph can be built from the output derivation *during* equation generation. This process is described in Algorithm 1.

4.2 Prompt Generation

Fine-tuning prompts. To assess mathematical derivations with a range of LLMs, we extract prompts following the template in the example below from the synthetic derivations:

$$\textit{Given} \quad q(a) = e^a$$

$$\textit{and} \quad G(a) = -e^a + \tfrac{d}{da}q(a),$$

$$\textit{then derive} \; -e^a + \tfrac{d}{da}q(a) = 0,$$

$$\textit{then obtain} \quad e^{G(a)} = 1$$

Premise equations are denoted by *Given* (or *and*), the result of evaluating integrals and derivatives is denoted by *then derive*, and the goal equation is denoted by *then obtain*. The prompt guides a derivation from premises, through certain intermediate steps, to the goal equation. These prompts are also used to evaluate fine-tuned models. The intermediate steps are integration or differentiation results intended to guide LLMs through the derivation (Wei et al., 2023) and reduce hallucinations.

Few-shot prompting GPT. Few-shot prompting (Møller et al., 2023), or in-context learning, is a method of enhancing the zero-shot performance of LLMs by including examples in the prompt. In particular, chain-of-thought prompts (Wei et al., 2023) involve explanations of how results have been ob-

tained from in-context examples, rather than simply including such results without explanation, and can improve generation quality. In our case, where the results themselves *are* equation chains serving as mathematical descriptions, any further chain-of-thought prompting would involve accompanying natural language explanations. Instead of including natural language, as this is something we aim to minimize, we create prompts that contain a variety of training examples that fit into a template.

For each static evaluation prompt such as the fine-tuning example, a set of 5 example prompts (and derivations) are randomly selected from the training set under the condition that *at least 2* training prompts contain "then derive" and "and", matching the earlier prompt example. This matching was chosen to emulate the training data, where a third of examples contain > 1 premise. The examples are then fit into the template:

The following examples consist of a prompt (denoted by Prompt:) and a mathematical derivation (denoted by Derivation:). Each derivation contains LaTeX equations separated by "and".

The training prompts are appended after this description, then the template continues:

Now given the following prompt, generate the derivation. Ensure equations are split by the word "and".

The evaluation prompt is inserted here, prepended by *"Prompt:"*.

This prompting methodology was chosen to minimize any natural language in the generated output, and to force derivations into the correct format (LaTeX equations split by "and"). Also, only the evaluation prompt is perturbed, ensuring consistency in the evaluation. This ensures that differences in the generated output are caused only by changes to the evaluation prompt. If in-context examples are not controlled for, then we would be measuring generation differences due to both the perturbation *and* changes to in-context examples. The instantiated few-shot template is fed to the GPT models through the OpenAI API[3], with temperature set to 0 to minimise non-deterministic effects.

4.3 Perturbations

A perturbation is a transformation applied to the input text and ground truth that ideally changes a single aspect of reasoning (*e.g.,* a change of notation, or swapping commutative expressions). We apply four perturbations to the static test set to evaluate generalisation, where each perturbation generates an additional test set later used for pairwise evaluation.

Variable Renaming. In the training set, derivations rely on a vocabulary of 155 symbols sampled from Wikipedia physics (Meadows et al., 2022) (*e.g.,* Ψ_{nl}, E_n, $\mathbf{J}_P$, η, g'_ε). For each example in the static set, we uniquely map each symbol to an out-of-distribution symbol sampled from 11 Greek letters (*e.g.,* $E_n = n + x$ becomes $\alpha = \beta + \gamma$).

Expression Exchange. In the training set, there is an asymmetry with respect to premises being defined with functions on the LHS and expressions on the RHS (*e.g.,* $E_n(n, x) = n + x$). However, operations are frequently used that can substitute LHS for RHS (and vice versa) in many cases, and both functions and operations may appear on either side of equations. Simply, we swap expressions either side of the equality for all equations in the static set (*e.g.,* $E_n = n + x$ becomes $n + x = E_n$).

Alternative Goal. In the training set, the goal equation in the prompt (denoted by *then obtain* in Section 3.2) is the final equation in the target text to be generated. For each example in the static set, we derive an alternative goal equation from the penultimate equation, by random selection of operators and operands. This perturbation should not incur large score differences between static and perturbed sets, because it is simply applying alternative in-distribution operations that occur frequently during training.

Step Removal. In the training set, equations that occur as a result of evaluating differentials and integrals are included in the prompt as intermediate steps. These are used to guide derivation generation (see Section 3.2). This perturbation removes such *"then derive"* equations from the prompt, which may have two effects: (1) where appropriate, a model may generate derivations of goal equations that *avoid* evaluating integrals etc., meaning that even valid derivations may have hugely different surface forms to target text; (2) where integration is necessary, the model will be forced to implicitly perform *multiple* operations in a single step. Although some equations will have been encountered during fine-tuning (*e.g.,* $\frac{d}{dx}\sin(x) = \cos(x)$), it is likely models will hallucinate results due to the imposed granularity gap forced by the perturbation at this step.

[3] https://platform.openai.com/overview

5 Data Analysis

The creation of the datasets involves initially generating annotated derivations, creating prompts from those derivations, splitting the data into training and static test sets, then perturbing the static set to form (four) perturbation sets. Table 1 describes the sizes of the various datasets. Discrepancies between perturbation set sizes arise mostly from the imposed 512 token limit, such that full reasoning chains are accepted by fine-tuned models. For example, if a derivation consists of nearly 512 tokens, then *Variable Renaming* is applied, the symbol change may extend the sequence beyond 512, so we exclude that example. The lower size of the *Step Removal* set simply comes the fact that not all derivations involve evaluating integrals or derivatives, and this perturbation applies only to those that do.

Dataset	Size
Training	15.3k
Static Test Set	3.1k
Variable Renaming	2.9k
Expression Exchange	3.1k
Alternative Goal	3.1k
Step Removal	1.0k

Table 1: Dataset sizes. *Static* is the held-out test set, while *Variable Renaming* and the other three perturbed sets are generated by applying perturbations to the static set.

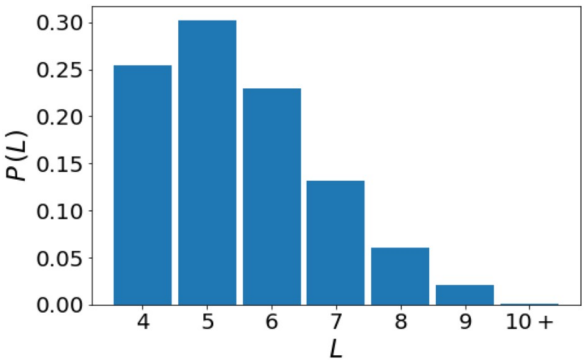

Figure 2: $P(L)$ is the probability that a given derivation features L equations.

Fig. 2 shows the distribution of derivation lengths in the training set. The peak at $L = 5$ arises from derivations initially being generated with lengths following a truncated Gaussian centered at $L = 7$ ($L > 3, \sigma = 3$). A large proportion were later excluded due to the token limit, reducing the distribution maximum. New equations are generated by applying one of *18 operations* to a previous equation in the sequence. For a random equation, the probability it was formed by applying operator O is given in Fig. 4. This distribution is implicitly controlled by hyperparameters. As mentioned in Section 4.3, there is an asymmetry between the LHS and RHS of the equations, the effects of which are explored by *Expression Exchange*. This is reflected in the relative probability between S_L and S_R, which are respectively substitution operations for the LHS and RHS. Also, $\int_E$ (evaluate integrals) is less common than ∂_E (evaluate derivatives) largely because many integrals evaluate to (excluded) piece-wise functions, and other long equations. We omit LaTeX equation strings longer than 350 characters.

A derivation may also be characterized by specific permutations (chains) of the set of operators that formed it. For a given length, the total permutations and related information is displayed in Table 5. Across all derivation lengths, the chain $\partial \rightarrow \partial_E \rightarrow S_L$ (and its integral equivalent) occurs particularly frequently, and contributes to longer chains. Decoding, this chain means a differential operator (∂) is applied to both sides of an equation ($\frac{d}{dx}y(x) = \frac{d}{dx}\sin(x)$), the evaluate derivatives operator (∂_E) is applied ($\frac{d}{dx}y(x) = \cos(x)$), then the LHS substitution operation (S_L) is applied ($\frac{d}{dx}\sin(x) = \cos(x)$).

	ROUGE	BLEU	BLEURT	GLEU
T5-base (f)	88.6	81.3	70.5	83.4
FLAN-T5-base (f)	87.3	79.4	68.9	81.7
T5-large (f)	89.4	82.8	72.1	84.7
FLAN-T5-large (f)	**90.2**	**84.6**	**73.2**	**86.1**
T5-base	89.5	82.8	70.5	84.4
FLAN-T5-base	87.0	80.3	67.2	81.9
T5-large	91.0	85.1	72.5	86.4
FLAN-T5-large	**91.2**	**86.1**	**72.9**	**87.2**
ChatGPT	80.3	70.8	63.1	73.5
GPT-4	82.8	72.2	62.9	75.6

Table 2: Static evaluation results. The four models associated with (f) are evaluated on a set of 2K examples. Remaining models are evaluated on 100 examples.

6 Empirical Evaluation

We first perform a conventional evaluation on the static set, then dynamically evaluate the effect of perturbations by comparing static scores with perturbed scores from the out-of-distribution test sets, following Meadows et al. (2023). Details of training and metrics (such as ROUGE) are given in

	ROUGE				BLEU				BLEURT				GLEU			
	VR	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**
T5-base (f)	80.2	86.2	88.3	77.0	74.8	78.2	80.9	64.2	67.7	67.3	67.4	51.5	76.1	80.4	83.1	69.2
FLAN-T5-base (f)	24.4	84.3	86.7	77.7	41.1	76.0	78.8	66.6	18.7	67.0	67.9	**56.8**	44.2	78.5	81.3	71.0
T5-large (f)	**85.0**	86.8	89.2	77.7	**79.3**	79.5	82.5	66.4	**70.8**	68.3	69.6	54.1	**80.8**	81.5	84.4	70.6
FLAN-T5-large (f)	83.0	**87.1**	**89.5**	**78.6**	78.5	**80.4**	**83.5**	**68.9**	69.0	**68.7**	**70.3**	56.1	79.6	**82.1**	**85.1**	**72.4**
T5-base	82.2	87.3	89.9	79.9	77.2	81.6	83.7	68.8	71.1	69.6	70.1	56.5	78.0	82.6	85.3	72.5
FLAN-T5-base	25.7	86.7	87.8	78.5	40.4	81.1	81.1	68.5	14.6	69.0	66.4	56.7	42.9	82.2	82.9	71.8
T5-large	**86.2**	87.7	**90.5**	80.6	**80.7**	82.4	84.7	71.0	**71.9**	**70.7**	**71.8**	59.6	**81.7**	83.3	86.1	74.1
FLAN-T5-large	85.1	**87.9**	90.4	**80.7**	79.8	**83.1**	**84.8**	**72.3**	71.2	70.5	71.4	**61.0**	80.6	**83.8**	**86.2**	**75.0**
ChatGPT	78.8	78.8	80.6	73.3	70.2	70.7	71.4	64.2	63.9	62.1	61.7	50.9	72.7	72.9	74.3	67.7
GPT-4	81.6	80.9	82.1	75.6	71.1	68.3	70.4	61.7	64.2	61.3	61.8	50.4	74.4	72.3	74.4	67.2

Table 3: Conventional scores on perturbation sets. Variable Renaming (**VR**) involves out-of-distribution symbol replacement. Expression Exchange (**EE**) involves swapping expressions either side of equality symbols. Alternative Goal (**AG**) involves using a different operation to derive an alternative goal equation. Step Removal (**SR**) involves removing intermediate steps from the prompt.

	ROUGE				BLEU				BLEURT				GLEU			
	VR	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**
T5-base (f)	8.4	**2.4**	0.3	11.6	6.5	**3.1**	0.4	17.1	2.8	3.2	3.1	19.0	7.3	**3.0**	0.3	14.2
FLAN-T5-base (f)	62.9	3.0	0.6	**9.6**	38.3	3.4	0.6	**12.8**	50.2	**1.9**	**1.0**	**12.1**	37.5	3.2	0.4	**10.7**
T5-large (f)	**4.4**	2.6	**0.2**	11.7	**3.5**	3.3	**0.3**	16.4	**1.3**	3.8	2.5	18.0	**3.9**	3.2	**0.3**	14.1
FLAN-T5-large (f)	7.2	3.1	0.7	11.6	6.1	4.2	1.1	15.7	4.2	4.5	2.9	17.1	6.5	4.0	1.0	13.7
T5-base	7.3	2.2	-0.4	9.6	5.6	1.2	-0.9	14.0	-0.6	**0.9**	0.4	14.0	6.4	1.8	-0.9	11.9
FLAN-T5-base	61.3	**0.3**	-0.8	8.5	39.9	-0.8	-0.8	11.8	52.6	-1.8	0.8	**10.5**	39.0	**-0.3**	-1.0	10.1
T5-large	4.8	3.3	0.5	10.4	4.4	2.7	**0.4**	14.1	**0.6**	1.8	0.7	12.9	4.7	3.1	**0.3**	12.3
FLAN-T5-large	6.1	3.3	0.8	10.5	6.3	3.0	1.3	13.8	1.7	2.4	1.5	11.9	6.6	3.4	1.0	12.2
ChatGPT	1.5	1.5	**-0.3**	**7.0**	**0.6**	**0.1**	-0.6	**6.6**	-0.8	1.0	1.4	12.2	**0.8**	0.6	-0.8	**5.8**
GPT-4	**1.2**	1.9	0.7	7.2	1.1	3.9	1.8	10.5	-1.3	1.6	**1.1**	12.5	1.2	3.3	1.2	8.4

Table 4: Model generalisability scores for performance. A higher absolute value implies a perturbation has a significant effect on the generation. A positive value implies the perturbation increased the problem difficulty, while a negative value implies the opposite. Bold values represent lowest absolute difference from static scores, where the perturbation has had the smallest effect on the sample score.

Appendix B.

6.1 In-distribution Derivations

Table 2 describes the relative performance of four fine-tuned models, and the GPT models evaluated using the few-shot prompt. The fine-tuned models are the base and large versions of T5 (Raffel et al., 2020) and FLAN-T5 (Chung et al., 2022). On 2K examples from the static set (denoted by (f)), FLAN-T5-large outperforms all models in all metrics. This notable advantage over T5-large is likely due to instruction fine-tuning – our prompt is an instruction. However, this advantage over T5 does not extend to FLAN-T5-base, which scores lower than T5-base in all metrics. This potentially due to fine-tuning instability observed in T5 (Asai et al., 2022). In addition to FLAN-T5-large (`MathT5-large`), fine-tuned T5-base (`MathT5-base`) is available online[4].

The other scores reported in Table 2 (without (f)) are evaluated on a set of 100 examples, sampled from the static set, *that contain integration and differentiation results in the prompt*. The lower sample size is because inference with GPT through the OpenAI API is financially inefficient, and there are four perturbed prompts for each prompt in the static set. The fine-tuned models score within 3 units of their previous scores, model rankings are preserved across all metrics, and we assume the GPT scores would report similarly for larger samples.

The (ROUGE) scores obtained by fine-tuned

[4] https://huggingface.co/jmeadows17/MathT5-base

FLAN-T5-large (91), T5-large (91), T5-base (89), and FLAN-T5-base (87) are all higher than the GPT models. *FLAN-T5-base is the worst fine-tuned model, but scores 4 ROUGE higher than GPT-4* (83), and 7 ROUGE higher than ChatGPT (80). This reflects that the general pre-training of the GPT models *does not capture granular reasoning with equations at the operational level*, compared to the fine-tuned models (as elicited through the few-shot prompt).

6.2 Perturbed Derivations

Table 3 shows the larger fine-tuned models outperform the GPT models on every test set, in every metric, on all perturbed test sets. However, we measure the impact of perturbations (and hence a model's generalisability) differently. Given static ground truth derivation s_i, model derivation $\hat{s}_i$, perturbed ground truth p_i, perturbed model derivation $\hat{p}_i$, and some metric M, the performance decrease due to a perturbation on the i^{th} derivation is given by $M(s_i, \hat{s}_i) - M(p_i, \hat{p}_i)$. It is this value, averaged over derivation pairs, that is reflected in Table 4. We aim to use this score difference to better determine the effect of perturbations, and hence better determine models' robustness and generalisation.

Of the fine-tuned models evaluated on the larger perturbed sets (f), according to ROUGE, the T5 models generalise worse than the FLAN-T5 models to all perturbations other than SR. According to BLEU, the T5 models both generalise better to EE and AG perturbations. According to BLEURT, the T5 models both generalise better to only VR. According to GLEU, they only generalise better to AG. This disparity in rankings and model scores makes it difficult to accurately determine which models generalise better to a given perturbation. However, it is important to note that they do mostly agree on the ranking of perturbation difficulty. Other than BLEURT, from least impactful to most, the perturbations are ranked AG (alternatve goal), EE (expression exchange), VR (variable renaming), SR (step removal). In particular, this suggests that SR and VR lead to derivations that heavily diverge from the target reference. The metrics also agree that FLAN-T5-base completely fails to generalise to out-of-vocabulary symbols (VR) (potentially due to fine-tuning instability).

Scores consistently report that the GPT models generalise to the VR and EE perturbations better than any of the fine-tuned models. GPT models generalise to all perturbations except those where critical results have been removed from the prompt (SR). This is also true for fine-tuned models, but they also generalise worse to perturbations affecting out-of-distribution structural symmetry (EE), and worse still, out-of-vocabulary symbols (VR). All models can derive alternative final equations of derivations that rely on a different final operation (AG). For VR, GPT's advantage stems from its training involving immense vocabularies, and that the few-shot prompt is designed to not bias any particular symbols. GPT's advantage in EE stems from the fine-tuned models training involving a symmetry bias (evidenced by Fig. 4 and Tab. 5), so breaking this symmetry, as EE does, results in lower scores. This symmetry bias *is* included in the few-shot prompt, but GPT is insensitive to this perturbation. This is likely due to the vast amount of mathematical examples it has seen during training, spanning equations and derivations with other symmetry biases.

As an alternative approach to measuring the effect of perturbations, we calculate the ratio $M(\hat{s}_i, \hat{p}_i)/M(s_i, p_i)$ that measures how much a model's perturbed derivation deviates from its static derivation in comparison to the ground truth pair. Generally, the scores are < 1, which indicates *the difference between static and perturbed derivations produced by models is less different than equivalent ground truth pairs*. Ratios averaged over the sample are given in Table 8, which highlights that BLEURT is comparing derivations in a way other metrics are not, and returns very high and low average ratios. Also, Fig. 3 shows the ROUGE distribution over the 100 samples, and displays behaviour not captured by either table. Table 4 shows that the GPT models feature VR and SR derivations that score *below 0.4 ROUGE*, whereas the distribution of each is narrower for each perturbation, for FLAN-T5-large. This suggests that GPT models may score reasonable average results, and they appear to generalise well at scale, but edge cases may negatively effect them more than specialised models.

7 Qualitative Analysis

7.1 Mathematical Proficiency in Static Derivations

In Appendix C.1, we take 12 derivations where FLAN-T5-large obtains around 91 ROUGE (its average performance), and compare its output with

that of ChatGPT and GPT-4 on the same derivations. In Appendix C.2 we randomly sample 13 derivations and compare each model's solution.

Higher scores are not faithful representations of derivation quality. While there is a positive correlation between scores and reasoning quality, a significant proportion of derivations with lower scores are actually of higher quality. We define a "high quality" derivation to include no incorrect, inconsistent, repeated, redundant, or irrelevant equations, with minimum skipped steps. Given this definition, it is observable how a high quality derivation may score poorly given the ground truth reference. A model's derivation may include alternative but correct reasoning, leading to significant divergence from the reference surface form. Another derivation might take the same reasoning path as the ground truth, but writes an incorrect equation (*e.g.,* omitting an exponent) leading to a higher score. This semantic error is not captured by the metrics.

Equations and syntax are rarely incorrect. Of the 75 static derivations analysed (3 model outputs per ground truth sample), only 5 equations were incorrect. FLAN-T5 omitted function names from premises written in the prompt, while ChatGPT made more semantic errors such as incorrect differentiation (Derivation 3) and integration (Derivation 23). GPT-4 generated all equations correctly. After minor post-processing for the purpose of rendering derivations in Appendices C and D (such as removing spaces between "\\" and *e.g.*"nabla" for FLAN-T5 due to new vocabulary), we detected only 7 syntax errors of the 1.7K equations rendered. GPT-4 did not contribute to any syntax error.

Models may deliver more succinct derivations comparative to ground truth references. Due to the randomness involved in applying operations to derive ground truth derivations (Table 5), it is possible that more efficient derivations exist. Models frequently find such derivations yet are penalized due to surface form divergence from the ground truth. Derivation 16 shows FLAN-T5 using a more efficient substitution.

Models apply composite operators or skip steps. Each model must write mathematically valid paths between equations defined in the prompt. These are either premise equations, integration or differentiation results, or the goal equation. It is most common for models to skip the penultimate equation in the derivation. This occurs either from models performing multiple operations in a single step, or from skipping steps entirely (Derivation 6).

GPT models derive irrelevant equations. GPT models (most prominently ChatGPT) tend to write irrelevant equations that (although not necessarily incorrect) correspond to an inefficient application of operators. Derivation 14 shows FLAN-T5 selecting an efficient path, but skipping the penultimate step. In comparison, GPT-4 unnecessarily writes (304) instead of evaluating just the integral. ChatGPT writes an incorrect equation (310), and uses it to form (311) which is never used.

GPT models write unsimplified equations. We fine-tune FLAN-T5-large on computer algebra derivations that automatically evaluate terms such as $x + x$ to $2x$ before outputting equations. ChatGPT (339) and GPT-4 (260) write equations that are unsimplified in this way, that are more unnatural, as they do not benefit from this training.

ChatGPT unnecessarily rearranges equations in convoluted forms. There are instances where ChatGPT has specifically used log rules (177, 270) to rearrange equations. It also attempts to rearrange in terms of square roots (510), among other oddities. This demonstrates that any results derived from ChatGPT may be written in some convoluted form.

ChatGPT diverges from the prompt. Some equations are defined in the prompt. Although models have to reason at the *inter*-equational level when ordering prompt equations in the derivation, at the *intra*-equational level there should be no difficulty simply stating the equation as is. Despite this, ChatGPT can rearrange prompt equations (404). Also, all necessary int/diff results are given in the prompt, yet ChatGPT unnecessarily decides to evaluate other derivatives or integrals.

Overall comparison. FLAN-T5-large produces higher quality derivations, on average, than the larger GPT models (Fig. 3), with a derivation style matching the computer algebra system. GPT models behave differently to this, but GPT-4 possesses a sharpness and coherence that ChatGPT lacks.

7.2 Perturbation Sensitivity and Generalisability

The static qualitative analysis reveals that metrics miss significant semantic information, even for single examples. We similarly analyse the effect of perturbations applied to the 13 static derivations from Appendix C.2 (230 perturbed derivations between the models). We note that conclusions drawn

from this section are limited by the size of the derivation sample. We avoid discussing conclusions already determined from the quantitative analysis for brevity.

7.2.1 Generalisation to Variable Renaming (VR)

The errors made by FLAN-T5 follow those of ChatGPT from the static analysis, namely, an increased rate of ***incorrect equations*** (Derivation 22 VR), longer derivations due to ***irrelevant steps*** (Derivation 21 VR) or ***less efficient paths*** (Derivation 20 VR), ***repeated equations*** (Derivation 14 VR), and ***prompt divergence*** (Derivation 23 VR). Particularly, FLAN-T5 is not just rearranging but is *incorrectly writing prompt equations* (582, 707).

VR errors not present in static derivations. There are additional errors made by FLAN-T5 that are unique to VR. Such as ***hallucinating variables*** (631, 755) and ***omitting equality symbols from premises*** (822). Note that GPT-4 makes *none* of these errors[5] yet scores 2 ROUGE less. ChatGPT also writes ***redundant equations*** (621, 622).

7.2.2 Generalisation to Expression Exchange (EE)

ChatGPT appears sensitive to expression exchange. ChatGPT makes similar errors as in the static analysis, but more frequently. Most notably, the number of incorrect equations *doubles* comparative to the static evaluation, ChatGPT repeats equations (986, 1022), and multiple equality signs have been used (1024). This is contrary to what the metrics report in Tab. 4.

EE errors not present in static derivations. FLAN-T5 begins to make a new category of error involving ***redundant equations*** such as $x = x$ (1010, 1033, 1058). Although not an error, a premise equation defined in the prompt (1102) is written with swapped LHS/RHS (*i.e.,* EE). GPT-4 writes a different integration constant to that given in the prompt (940).

7.2.3 Generalisation to Alternative Goal (AG)

ChatGPT appears sensitive to AG. In addition to the usual static errors, ChatGPT *doubles* the number of incorrect equations written. This is surprising given that this perturbation simply uses an *alternative final operation* to obtain the final equation, and is hence a very minor perturbation. Table 4 shows ROUGE differences of < 1 across all models.

AG errors not present in static derivations. "Derivation 18 AG" is uniquely incorrect as it is the only instance that *limits of integration* (1247) occur in any equation. Additionally, ***prompt divergence*** occurs as it does with GPT-4 – with ***alternative constants of integration*** (1223, 1224, 1245, 1246). Fig. 4 shows there is a clear *operation preference*. To derive an alternative goal equation, more unique alternative equations are obtained using *less likely combinations of operators* (Table 5). This affects GPT models through the randomised few-shot prompt. GPT-4 and ChatGPT independently select the same integration constant in the same derivation (Derivation 18 AG), which features the *only double integral* in the examined derivations. FLAN-T5 has been fine-tuned to only include integration constants from the training vocabulary, and does not make this change.

7.2.4 Generalisation to Step Removal (SR)

Models postpone integration or differentiation. SR removes diff/int results from the prompt. These results simultaneously guide the derivation through intermediate equations, while providing answers to difficult derivatives/integrals. In Section 3.3 we discussed two scenarios for SR derivations: **(1)** avoiding such results where appropriate; **(2)** performing the necessary evaluation with an increased likelihood of hallucination. In some SR derivations (14, 15, 21, 25) such evaluation is avoided, *or postponed until the final equation*, where it is given indirectly through the goal equation in the prompt. This postponement is an unexpected approach to **(1)**. In others (16, 18, 24), diff/int evaluation is performed in roughly the original place, following **(2)**. GPT-4 incorrectly integrates for the first time by a factor of -1 (1583).

SR errors not present in static derivations. Present in other perturbed derivations, ***redundant equations*** (*e.g.,* $x = x$) are now written by GPT-4 (1493, 1547) alongside the other models. Some of these are final equations dually resulting in a unique form of ***prompt divergence***. ChatGPT writes an "equation" (1517) that ***includes natural language***.

8 Correlating Metrics with Manual Scores

We have highlighted that metrics do not agree on various aspects, and more obviously that they do not account for mathematical errors. Here, we

[5] To make quick comparisons simply CTRL+F "Derivation 13 VR" then omit the "VR" (for derivations 13-25 inclusive).

further explore their ability to rank models and perturbations, and train a regression model to improve correlation with manual scores. We sample 100 static and perturbed derivations from Appendices C.2 and D, define a number of error categories based on the previous section, and formulate various manual rankings based on error counts within this sample. Rankings from generation metrics neither agree nor correlate with manual rankings, with respect to both *model performance* (Tab. 9) and whether a model generalises better to a given perturbation than another model (Tab. 10). However, they agree and correlate better with rankings based on generalisation to each perturbation (Tab. 11).
In particular, we use the error categories to define a scoring function (separately to the manual rankings), and train an XGBoost regressor (Chen and Guestrin, 2016) to learn this function using various scores from other metrics. We define fixed weights $\mathbf{w}$, that determine the importance of categories, and obtain scores for each derivation through the following function:

$$M(\mathbf{x}; \mathbf{w}, \alpha) = \alpha\left(\exp\left\{ \ln\left(\frac{\alpha+1}{\alpha} \right) \mathbf{w} \cdot \mathbf{x} \right\} - 1 \right)$$

This heavily penalizes *incorrect* and *irrelevant equations*, and *overall incorrectness*. The exploratory XGB metric returns scores that naturally correlate far better with the scoring function, over the derivation sample, than the generation metrics (Tab. 12), and it leads to improved correlation coefficients for model performance and generalisation rankings, but *lower* coefficients for perturbation generalisation rankings. This demonstrates that correlation may be improved and, combined with correlation coefficients from the generation metrics, suggests a potential trade-off between what current metrics are able to effectively rank. No metric scores positive coefficients in all three rankings. Appendix A describes the above in more detail.

9 Conclusion

In this work, we use a derivation generation algorithm to create data for fine-tuning and evaluating LLMs, involving granular multi-step mathematical reasoning with equations. We fine-tune and prompt a range of LLMs to generate LaTeX derivations, and systematically perturb prompts and derivations to determine model generalisability and out-of-distribution mathematical proficiency. This reveals critical insights for mathematical text generation and reasoning with LLMs.

Fine-tuning with synthetic reasoning data can lead to performance advantages over larger models. The parameter count of FLAN-T5-large is respectively 2 and (reportedly) 6 orders of magnitude smaller than ChatGPT and GPT-4. Yet, after fine-tuning on synthetic derivations, its performance surpasses the few-shot performance of ChatGPT and GPT-4 in conventional scores, even on out-of-distribution examples, in every metric. This suggests that the careful fine-tuning of open source foundation models (Rombach et al., 2021; Touvron et al., 2023) may lead to more effective inference in specialized fields, such as mathematics or physics.
Text generation metrics fail to capture fine-grained semantic failures of intricate reasoning. Without suitable metrics it is challenging to determine if generated reasoning is sound. In mathematical language, there is a high sensitivity to perturbations as a consequence of the modality's stricter reliance on structure and long-range dependencies. A minute change to a bracket or exponent can have devastating effects on the mathematical validity of a statement. The canonical metrics explored in this work *entirely* fail to capture this semantic nuance, *even failing to correctly rank model performance* in multiple categories. Moreover, the out-of-distribution performance of models is revealed only by considering the score difference between static and perturbed examples, rather than conventional scores on perturbed sets alone. Experiments training a regressor (XGBoost) to approximate a custom scoring function generally leads to better correlation with manual rankings. However, the regressor fails where other metrics succeed, highlighting a trade-off in ranking proficiency. A suitable metric should overcome such trade-offs, while delivering a fair score of the validity of individual derivations, ideally without relying on references (Ke et al., 2022). Synthetic derivations may be used to enhance the mathematical capabilities of LLMs, but it is of critical importance to develop suitable metrics to support LLM-based mathematical reasoning in specialized fields.

10 Limitations

Post-processing. Additional tokens were necessarily added to the T5 vocabulary that are space separated from other tokens during inference (such as "\\"). Also, "\\" is omitted in cases where it should be the first instance of a sequence. These are issues when trying to render LaTeX. We do not

correct this when evaluating output derivations in the quantitative analysis (it would make little difference if we did), but we do correct it to render the 1.7K equations in the Appendix. On the Hugging Face repository we include a pipeline allowing the `MathT5` models to be used for inference with the hyperparameters used in this paper, which includes this post-processing.

Few-shot prompts. We have not exhaustively tested available prompts including chain-of-thought prompting and others. We have however ensured that static prompts reliably output derivations that match the specific template we fit derivations into; a sequence of solely equations (without "$") separated by "and". In over 50 static derivations tested, there were no examples that diverged from the template. However, it is difficult to assess whether a given prompt is optimal without repetitively evaluating a large number of samples, which, given the OpenAI API pricing, can get expensive.

Fine-tuning. We have also not exhaustively tested training hyperparameters. It is likely that the performance of the fine-tuned models could be improved further.

Synthetic derivations. The synthetic derivations are designed to capture highly granular reasoning at the operational level. Compared to real-world workings this is overly verbose. Further, while we have explored the most frequent operation chains (Table 5) that characterize derivations, many of these combinations are those which would rarely be employed in practice. To prompt a human to derive equations as the fine-tuned models do, one would ask them to write an equation for *each* operation they use, many of which would be taken for granted. However, the fine-tuned models also skip steps.

References

Toufique Ahmed and Premkumar Devanbu. 2022. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–5.

Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. 2022. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6655–6672, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Ernest Davis. 2019. Proof verification technology and elementary physics. In *Algorithms and Complexity in Mathematics, Epistemology, and Science: Proceedings of 2015 and 2016 ACMES Conferences*, pages 81–132. Springer.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32):e2123433119.

Deborah Ferreira and André Freitas. 2020. Natural language premise selection: Finding supporting statements for mathematical text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2175–2182.

Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, Julia Rozanova, and Andre Freitas. 2022. To be or not to be an integer? encoding variables for mathematical text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 938–948.

Emily First, Markus N Rabe, Talia Ringer, and Yuriy Brun. 2023. Baldur: Whole-proof generation and repair with large language models. *arXiv preprint arXiv:2303.04910*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*.

Naveen Sundar Govindarajalulu, Selmer Bringsjord, and Joshua Taylor. 2015. Proof verification and proof discovery for relativity. *Synthese*, 192:2077–2094.

Konstantin Hebenstreit, Robert Praas, Louis P Kiesewetter, and Matthias Samwald. 2023. An automatically discovered chain-of-thought prompt generalizes to novel models and datasets. *arXiv preprint arXiv:2305.02897*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Geoffrey C Hulette, Robert C Armstrong, Jackson R Mayo, and Joseph R Ruthruff. 2015. Theorem-proving analysis of digital control logic interacting with continuous dynamics. *Electronic Notes in Theoretical Computer Science*, 317:71–83.

Albert Qiaochu Jiang, Wenda Li, Jesse Michael Han, and Yuhuai Wu Lisa. 2021. Language models of isabelle proofs. In *6th Conference on Artificial Intelligence and Theorem Proving*.

Cezary Kaliszyk, Josef Urban, Umair Siddique, Sanaz Khan-Afshar, Cvetan Dunchev, and Sofiene Tahar. 2015. Formalizing physics: automation, presentation and foundation issues. In *International Conference on Intelligent Computer Mathematics*, pages 288–295. Springer.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Charlie-Ray Mann, Thomas J Sturges, Guillaume Weick, William L Barnes, and Eros Mariani. 2018. Manipulating type-i and type-ii dirac polaritons in cavity-embedded honeycomb metasurfaces. *Nature communications*, 9(1):1–11.

Jordan Meadows and André Freitas. 2021. Similarity-based equational inference in physics. *Physical Review Research*, 3(4):L042010.

Jordan Meadows and Andre Freitas. 2022. A survey in mathematical language processing. *arXiv preprint arXiv:2205.15231*.

Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023. A symbolic framework for systematic evaluation of mathematical reasoning with transformers.

Jordan Meadows, Zili Zhou, and Andre Freitas. 2022. Physnlu: A language resource for evaluating natural language understanding and explanation coherence in physics. *arXiv preprint arXiv:2201.04275*.

Aaron Meurer, Christopher P Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K Moore, Sartaj Singh, et al. 2017. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David A Plaisted. 1993. Equational reasoning and term rewriting systems. *Handbook of logic in artificial intelligence and logic programming*, 1:273–364.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*.

Varot Premtoon, James Koppel, and Armando Solar-Lezama. 2020. Semantic code search via equational reasoning. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 1066–1082.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Marco Valentino, Deborah Ferreira, Mokanarangan Thayaparan, André Freitas, and Dmitry Ustalov. 2022. Textgraphs 2022 shared task on natural language premise selection. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 105–113.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language. *arXiv preprint arXiv:2104.01112*.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. *arXiv preprint arXiv:2205.12910*.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models.

Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and Andre Freitas. 2023. Large language models, scientific knowledge and factuality: A systematic analysis in antibiotic discovery. *arXiv preprint arXiv:2305.17819*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.

A Correlating Metrics with Manual Scores

The qualitative analysis shows that that metrics are failing to capture important error types causing deviations from reference derivations, such as incorrect or irrelevant equations. They also penalize derivations that differ by taking correct but alternative reasoning paths. As we briefly discuss in the quantitative analysis, they fail to agree on what models generalise better to perturbations. At the same time, they largely agree on the difficulties of the perturbations themselves, in the context of individual models. In this section, we further explore their ability to rank models and perturbations, and train a regression model to improve some correlation scores at the expense of others.

Learning a manual scoring function. To construct a scoring function, we first sample 100 random derivations from Appendices C.2 and D, evenly distributed between ChatGPT and FLAN-T5-large. The set is evenly distributed between VR, EE, AG, and SR, and static derivations are overrepresented. We define 6 categories for the sample derivations: **overall** determines if the entire derivation is incorrect overall, **skip** indicates the presence of any skipped steps, **repeat** denotes the presence of repeated equations, **incorrect** denotes the presence of clearly incorrect equations, **irrelevant** denotes if equations are derived but do not

contribute to the goal equation, **redundant** equations denotes the existence of equations where the RHS is too obviously equal to the LHS, such as $\cos(x)\sin(x) = \sin(x)\cos(x)$ (1550). Each generated derivation is associated with a one-hot encoding $\mathbf{x}$, that represents the 6 categories. We define fixed weights $\mathbf{w}$, that determine the importance of categories in some idealized score. We obtain manual scores for each derivation through the following function

$$M(\mathbf{x}; \mathbf{w}, \alpha) = \alpha\left(\exp\left\{\ln\left(\frac{\alpha+1}{\alpha}\right)\mathbf{w}\cdot\mathbf{x}\right\} - 1\right),$$

where $\mathbf{w} = [w_{\text{ovr}}, w_{\text{skip}}, w_{\text{rep}}, w_{\text{inc}}, w_{\text{irr}}, w_{\text{red}}] = [0.2, 0.05, 0.15, 0.25, 0.25, 0.1]$, and $\alpha = 0.001$. This weighting scheme heavily penalizes *incorrect* and *irrelevant equations*, and *overall incorrectness*. The category value x_{ovr} is somewhat subjective but highly correlates with x_{inc}. Therefore, despite overall correctness being (arguably) the most important factor, w_{ovr} is lower due to the high w_{inc}. Otherwise, this choice of function (and α) ensures derivations score 1 without errors, 0 with all error types, and strong penalization of any error. Score bunching at $M(\mathbf{x}; \mathbf{w}, \alpha) < 0.5$ means a derivation might not contain an error from each category, yet may still be comparably wrong.

We fit an XGBoost regressor (Chen and Guestrin, 2016) to the derivation sample. We train with 8 features, including conventional ROUGE, BLEU, BLEURT, and GLEU scores, and the corresponding perturbation ratios described in Table 8. This ratio is set to 1 for static derivations. XGBoost learns the scoring function with a correlation of 1.0000 (Table 12). These scores are denoted in Tables 7-9 as **XGB**. We avoid any claim that this metric is suitable for this task, nor math generation tasks in general. It is considered for the purpose of exploring the possibility of accounting for error types that other metrics miss.

Formulating rankings. A suitable metric should at least form correct rankings of aspects such as model performance and generalisability with respect to important semantic features of generated text. To reduce subjectivity from the defined manual scoring function, we rely on the more objective *incorrect*, *irrelevant*, and *redundant equation* categories to guide manual rankings described in Tables 6 and 7. We consider three ranking schemes: ***model performance rankings*** on static and perturbed test sets as measured by conventional metric scores (Table 9), ***model generalisation rankings*** on perturbed sets calculated from the difference between static and perturbed scores (Table 10), and ***perturbation generalisation rankings*** that do not compare models, but rather rank perturbations with respect to how well each model generalises to them (Table 11).

Ranking proficiency trade-off. The results echo the quantitative analysis on much larger test sets. Namely, metrics are failing to rank model performance based on conventional *e.g.* ROUGE scores alone (Tab. 9). This is a major problem because one might not have access to reasoning examples before a perturbation was made, or may be evaluating on other out-of-distribution examples, and they would be unable to reliably determine which model is better. Even if one *does* have access both to static and perturbed scores (Tab. 10) this difficulty persists, as in Tab. 4. However, it is reconfirmed from the quantitative analysis (Tab. 11) that, at least within a single model's scope, metrics agree with manual rankings on perturbation difficulty (notably GLEU). Collectively, this suggests some trade-off between what metrics can proficiently rank. ***Not a single metric obtains a positive correlation coefficient in all three rankings.*** The introduction of the learned XGB metric reinforces this trade-off. It scores higher in model performance and generalisation rankings, but fails to rank perturbations. It also gives the lowest static scores (Tab. 6) and the highest score difference due to perturbations (Tab. 7).

B Training Details and Metrics

We fine-tune the T5 and FLAN-T5 transformer models (base + large) for 25 epochs, with a learning rate of 5e-5, a batch size of 4, and a random seed of 42, using the Adafactor optimizer. We run a validation epoch after each training epoch, using the ROUGE-2 score as the performance metric with early stopping. Additionally, the model vocabulary is extended to better fit our custom dataset. Token embeddings in the model are resized to maintain consistency with the updated vocabulary. Of the metrics, ROUGE (Lin, 2004) is used for evaluating automatic summarization and machine translation, and measures the quality of a summary by comparing it to reference summaries using n-gram overlap. BLEU (Papineni et al., 2002) is used similarly and considers what proportion of the generated words are present in the reference text, including a penalty for sentences that are too short. BLEURT (Sellam

Length	Permutations	Chain	P(Chain)	Relative Frequency
4	842	$\partial \rightarrow \partial_E \rightarrow S_L$	0.0369	31
		$\int \rightarrow \int_E \rightarrow S_L$	0.0186	16
5	2850	$+ \rightarrow \partial \rightarrow \partial_E \rightarrow S_L$	0.0053	15
		$- \rightarrow \partial \rightarrow \partial_E \rightarrow S_L$	0.0048	14
6	3163	$\int \rightarrow \int_E \rightarrow S_L \rightarrow \int_E \rightarrow S_R$	0.0033	11
		$+ \rightarrow \partial \rightarrow \partial_E \rightarrow S_L \rightarrow S_R$	0.0020	6
7	2081	$\times \rightarrow \div \rightarrow R \rightarrow S_R \rightarrow S_R \rightarrow X^O$	0.0009	2
		$\div \rightarrow \int \rightarrow \partial \rightarrow \int_E \rightarrow S_L \rightarrow \partial_E$	0.0009	2

Table 5: For a given derivation length, *Permutations* describes the number of unique operation sequences present in the training data. *Chain* describes the two most frequent operation sequences based on symbols defined in Fig. 4. *P(Chain)* is the probability of the chain, and *Relative Frequency* is its relative probability of occurrence compared to the average.

Perturbation	Model	Incorrect eqs	Repeating eqs	Redundant eqs	Manual rank	ROUGE	BLEU	BLEURT	GLEU	XGB
Static	FLAN-T5	2	0	0	2	92	86	76	87	77
	GPT-4	0	0	0	1	84	73	65	76	55
	ChatGPT	2	2	0	3	79	69	64	72	47
VR	FLAN-T5	6	2	1	3	85	77	75	78	55
	GPT-4	0	0	0	1	85	77	70	80	53
	ChatGPT	4	1	1	2	81	75	70	76	41
EE	FLAN-T5	5	3	2	3	90	85	73	85	19
	GPT-4	0	0	0	1	84	75	65	78	34
	ChatGPT	4	2	0	2	82	75	62	77	26
AG	FLAN-T5	3	1	0	2	94	90	79	91	56
	GPT-4	0	0	0	1	84	73	66	77	34
	ChatGPT	4	1	0	3	84	78	70	79	22
SR	FLAN-T5	3	1	1	2	85	78	65	79	43
	GPT-4	3	0	2	1	79	66	54	70	38
	ChatGPT	5	0	2	3	76	66	55	69	26

Table 6: Values used for calculating Spearman's rank coefficients in Table 9. Rounded values are not used to make calculations, and are for ease of reading. Incorrect/repeating/redundant eqs represents the number of derivations that include that error (not the total number of errors). A model with highest performance has a manual rank of 1.

Model	Perturbation	Incorrect eqs	Repeating eqs	Redundant eqs	Manual rank	ROUGE	BLEU	BLEURT	GLEU	XGB
FLAN-T5	VR	4	2	1	4	6.1	8.3	1.5	8.3	22.0
	EE	3	3	2	3	1.0	-0.1	2.8	0.3	57.2
	AG	1	1	0	1	-0.4	-1.5	-0.5	-1.3	20.6
	SR	1	1	1	2	7.9	9.3	12.1	8.3	33.5
GPT-4	VR	0	0	0	1	0.8	0.3	-3.6	0.4	2.2
	EE	0	0	0	1	1.6	3.5	0.6	3.1	20.7
	AG	0	0	0	1	1.6	4.4	0.5	3.2	20.8
	SR	3	0	2	4	5.8	8.0	11.5	6.3	17.4
ChatGPT	VR	2	-1	1	2	-0.2	-3.4	-8.2	-2.3	6.2
	EE	2	0	0	3	0.9	-0.5	1.5	-0.5	21.2
	AG	2	-1	0	1	1.1	-0.5	0.6	-0.5	25.0
	SR	3	-2	2	4	5.4	5.0	11.1	3.9	21.3

Table 7: Values used for calculating Spearman's rank coefficients in Tables 10 and 11. Incorrect/repeating/redundant eqs represents the number of derivations that include that error in the relevant perturbed derivations, *minus* those from the static derivations. Exactly these values are used in calculations. A model generalises best to a perturbation with a manual rank of 1.

et al., 2020) is a learned evaluation metric trained on human-annotated data, accounts for complex linguistic phenomena, and correlates well with human judgement. GLEU (Mutton et al., 2007) is a

	ROUGE†				BLEU†				BLEURT†				GLEU†			
	VR	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**	**VR**	**EE**	**AG**	**SR**
T5-base	0.88	0.90	0.93	0.82	0.93	0.84	0.90	0.75	1.72	0.92	0.93	0.77	0.91	0.84	0.90	0.77
FLAN-T5-base	0.71	0.88	0.92	0.81	0.71	0.81	0.88	0.73	0.70	0.91	0.85	0.73	0.73	0.81	0.89	0.76
T5-large	0.94	0.90	0.93	0.82	0.96	0.85	0.90	0.75	1.03	0.92	0.92	0.79	0.95	0.85	0.91	0.77
FLAN-T5-large	0.91	0.91	0.93	0.82	0.95	0.85	0.89	0.75	2.08	0.92	0.92	0.79	0.94	0.85	0.90	0.77
ChatGPT	0.92	0.84	0.86	0.77	0.88	0.75	0.80	0.69	0.30	0.87	0.79	0.72	0.89	0.76	0.82	0.72
GPT-4	0.94	0.88	0.91	0.83	0.96	0.79	0.86	0.76	1.29	0.90	0.85	0.79	0.95	0.81	0.88	0.78

Table 8: Each value is given by $M(\hat{s},\hat{p})/M(s,p)$, where s and p are static and perturbed ground truth derivations, and $\hat{s}$ and $\hat{p}$ are corresponding predictions, for $M \in \{\text{ROUGE, BLEU, BLEURT, GLEU}\}$. A score < 1 indicates that, on average, the similarity between static and perturbed ground truths is higher than the similarity between predictions. A low score therefore means a given perturbation causes predictions that diverge from static predictions (normalized for the effect of that perturbation on the ground truth).

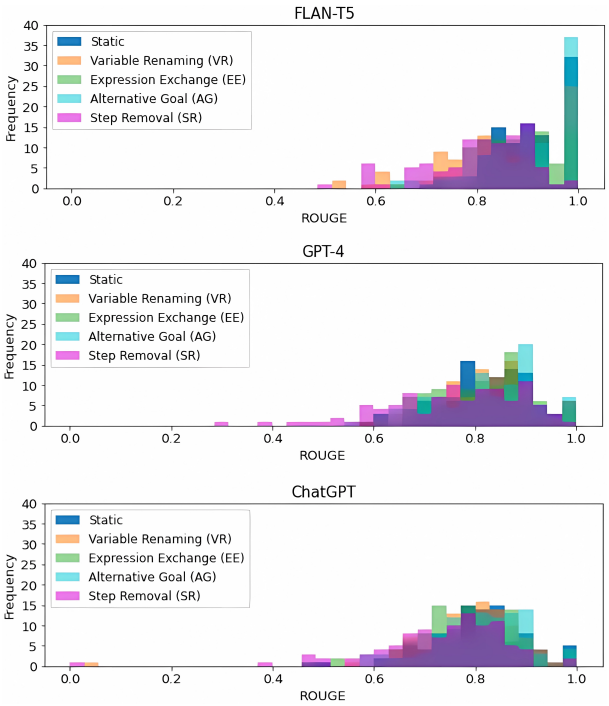

Figure 3: The ROUGE distribution over the sample of 100 derivations. *Static* (dark blue) represents unperturbed derivations, while other colors represent perturbed derivations.

	Static	VR	EE	AG	SR	Average
ROUGE	0.5	−0.5	−0.5	0.5	0.5	0.1
BLEU	0.5	−0.5	−0.5	−0.5	0.5	−0.1
BLEURT	0.5	−0.5	−0.5	−0.5	−0.5	−0.3
GLEU	0.5	0.5	−0.5	0.5	−0.5	0.1
XGB	0.5	−0.5	1.0	0.5	0.5	**0.4**

Table 9: **Model performance rankings.** Spearman's ranking coefficients calculated by comparing metric-based rankings to manual rankings, based on the conventional performance of FLAN-T5-large, GPT-4, and ChatGPT, from Table 6.

	VR	EE	AG	SR	Average
ROUGE	0.5	−0.5	−0.5	−0.5	−0.3
BLEU	0.5	−0.5	−0.5	−0.5	−0.3
BLEURT	0.5	1.0	0.5	−0.5	**0.4**
GLEU	0.5	−0.5	−0.5	−0.5	−0.3
XGB	1.0	1.0	0.5	−1.0	**0.4**

Table 10: **Model generalisation rankings**. Spearman coefficients show how accurately each metric ranks *each model's generalisation to a given perturbation*, measured by a performance drop from the static scores (such as Table 4), calculated from Table 7.

	FLAN-T5	GPT-4	ChatGPT	Average
ROUGE	0.4	0.8	0.4	0.5
BLEU	0.4	0.8	0.6	0.6
BLEURT	0.2	0.8	0.8	0.6
GLEU	0.6	0.8	0.6	**0.7**
XGB	0.4	−0.3	−0.2	0.0

Table 11: **Perturbation generalisation rankings.** Spearman coefficients show how accurately each metric ranks *model generalisation to each perturbation*, measured by a performance drop from the static scores, calculated from Table 7.

	Correlation
ROUGE	0.47
BLEU	0.31
BLEURT	0.34
GLEU	0.36
XGB	1.00

Table 12: Pearson correlation coefficients between metric scores and manual scores based on several categories of derivation errors.

variant of BLEU for evaluating dialogue systems, and measures n-gram overlap between generated and reference sentences, considering both precision and recall.

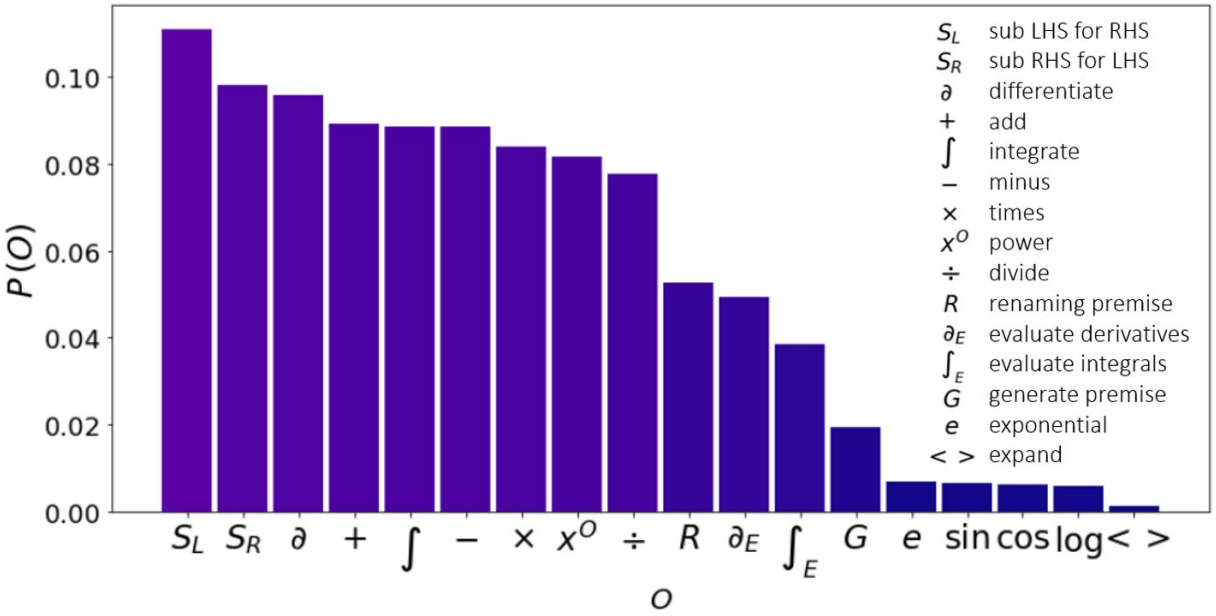

Figure 4: $P(O)$ is the probability that operation O is used to generate a given equation.

C Static analysis

C.1 Derivations where FLAN-T5-large obtains 91 ROUGE

C.1.1 Derivation 1

Correct

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{1}$$

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f \tag{2}$$

$$\frac{\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right)}{\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f} = 1 \tag{3}$$

$$\frac{\int (\mathbf{J}+\mathbf{v})d\mathbf{J}}{\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f} = 1 \tag{4}$$

FLAN-T5

91.0 88.9 70.7 89.0

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{5}$$

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f \tag{6}$$

$$\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f = \int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{7}$$

$$\frac{\int (\mathbf{J}+\mathbf{v})d\mathbf{J}}{\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f} = 1 \tag{8}$$

GPT-4

89.7 89.2 71.4 89.3

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{9}$$

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f \tag{10}$$

$$\int (\mathbf{J}+\mathbf{v})d\mathbf{J} = \frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f \tag{11}$$

$$\frac{\int (\mathbf{J}+\mathbf{v})d\mathbf{J}}{\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f} = 1 \tag{12}$$

ChatGPT

76.9 70.8 53.8 70.9

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{13}$$

$$\frac{\partial}{\partial \mathbf{J}}\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \frac{\partial}{\partial \mathbf{J}}\int (\mathbf{J}+\mathbf{v})d\mathbf{J} \tag{14}$$

$$\frac{\partial}{\partial \mathbf{J}}\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \mathbf{J}+\mathbf{v} \tag{15}$$

$$\mathrm{F_c}\left(\mathbf{J},\mathbf{v}\right) = \frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f \tag{16}$$

$$\frac{\int (\mathbf{J}+\mathbf{v})d\mathbf{J}}{\frac{\mathbf{J}^2}{2} + \mathbf{J}\mathbf{v} + f} = 1 \tag{17}$$

C.1.2 Derivation 2

Correct

$$f(C_d) = C_d \tag{18}$$

$$\frac{d}{dC_d} f(C_d) = \frac{d}{dC_d} C_d \tag{19}$$

$$\frac{d}{dC_d} f(C_d) = 1 \tag{20}$$

$$1 = \frac{1}{\frac{d}{dC_d} f(C_d)} \tag{21}$$

$$1 = \frac{1}{\frac{d}{dC_d} C_d} \tag{22}$$

$$1 = \frac{1}{\frac{d}{df(C_d)} f(C_d)} \tag{23}$$

FLAN-T5

91.5 90.4 71.4 90.5

$$f(C_d) = C_d \tag{24}$$

$$\frac{d}{dC_d} f(C_d) = \frac{d}{dC_d} C_d \tag{25}$$

$$\frac{d}{dC_d} f(C_d) = 1 \tag{26}$$

$$\frac{d}{dC_d} C_d = 1 \tag{27}$$

$$1 = \frac{1}{\frac{d}{dC_d} C_d} \tag{28}$$

$$1 = \frac{1}{\frac{d}{df(C_d)} f(C_d)} \tag{29}$$

GPT-4

92.3 94.4 72.0 94.4

$$f(C_d) = C_d \tag{30}$$

$$\frac{d}{dC_d} f(C_d) = \frac{d}{dC_d} C_d \tag{31}$$

$$\frac{d}{dC_d} f(C_d) = 1 \tag{32}$$

$$\frac{1}{\frac{d}{df(C_d)} f(C_d)} = \frac{1}{\frac{d}{dC_d} C_d} \tag{33}$$

$$1 = \frac{1}{\frac{d}{df(C_d)} f(C_d)} \tag{34}$$

ChatGPT

77.0 75.7 70.1 76.3

$$f(C_d) = C_d \tag{35}$$

$$\frac{d}{dC_d} f(C_d) = 1 \tag{36}$$

$$\frac{d}{df(C_d)} f(C_d) = \frac{d}{dC_d} f(C_d) \tag{37}$$

$$\frac{d}{df(C_d)} f(C_d) = 1 \tag{38}$$

$$1 = \frac{1}{\frac{d}{df(C_d)} f(C_d)} \tag{39}$$

C.1.3 Derivation 3

Correct

$$\mathrm{A_y}\left(\phi_2\right) = \cos\left(\phi_2\right) \tag{40}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = \frac{d}{d\phi_2}\cos\left(\phi_2\right) \tag{41}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{42}$$

$$\frac{d}{d\phi_2}\cos\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{43}$$

$$\frac{d^2}{d\phi_2^2}\cos\left(\phi_2\right) = \frac{d}{d\phi_2} - \sin\left(\phi_2\right) \tag{44}$$

$$\frac{d^3}{d\phi_2^3}\cos\left(\phi_2\right) = \frac{d^2}{d\phi_2^2} - \sin\left(\phi_2\right) \tag{45}$$

FLAN-T5

91.3 86.9 80.2 86.9

$$\mathrm{A_y}\left(\phi_2\right) = \cos\left(\phi_2\right) \tag{46}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = \frac{d}{d\phi_2}\cos\left(\phi_2\right) \tag{47}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{48}$$

$$\frac{d^2}{d\phi_2^2}\,\mathrm{A_y}\left(\phi_2\right) = \frac{d}{d\phi_2} - \sin\left(\phi_2\right) \tag{49}$$

$$\frac{d^2}{d\phi_2^2}\cos\left(\phi_2\right) = \frac{d^2}{d\phi_2^2} - \sin\left(\phi_2\right) \tag{50}$$

$$\frac{d^3}{d\phi_2^3}\cos\left(\phi_2\right) = \frac{d^2}{d\phi_2^2} - \sin\left(\phi_2\right) \tag{51}$$

GPT-4

89.8 81.2 71.8 82.6

$$\mathrm{A_y}\left(\phi_2\right) = \cos\left(\phi_2\right) \tag{52}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = \frac{d}{d\phi_2}\cos\left(\phi_2\right) \tag{53}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{54}$$

$$\frac{d^2}{d\phi_2^2}\cos\left(\phi_2\right) = -\cos\left(\phi_2\right) \tag{55}$$

$$\frac{d^3}{d\phi_2^3}\cos\left(\phi_2\right) = \frac{d^2}{d\phi_2^2} - \sin\left(\phi_2\right) \tag{56}$$

ChatGPT

75.7 56.7 55.0 63.6

$$\mathrm{A_y}\left(\phi_2\right) = \cos\left(\phi_2\right) \tag{57}$$

$$\frac{d}{d\phi_2}\,\mathrm{A_y}\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{58}$$

$$\frac{d^2}{d\phi_2^2}\cos\left(\phi_2\right) = -\sin\left(\phi_2\right) \tag{59}$$

$$\frac{d^3}{d\phi_2^3}\cos\left(\phi_2\right) = \frac{d^2}{d\phi_2^2} - \sin\left(\phi_2\right) \tag{60}$$

C.1.4 Derivation 4

Correct

$$l(M_E) = \cos(M_E) \tag{61}$$

$$\frac{d}{dM_E} l(M_E) = \frac{d}{dM_E} \cos(M_E) \tag{62}$$

$$\frac{d}{dM_E} l(M_E) - \frac{d}{dM_E} \cos(M_E) = 0 \tag{63}$$

$$\sin(M_E) + \frac{d}{dM_E} l(M_E) = 0 \tag{64}$$

$$\sin(M_E) + \frac{d}{dM_E} \cos(M_E) = 0 \tag{65}$$

$$\int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E = \int 0 dM_E \tag{66}$$

$$\int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E - 1 = \int 0 dM_E - 1 \tag{67}$$

$$y' - 1 = \int 0 dM_E - 1 \tag{68}$$

$$y' - 1 = \int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E - 1 \tag{69}$$

FLAN-T5

90.9 86.7 74.1 87.3

$$l(M_E) = \cos(M_E) \tag{70}$$

$$\frac{d}{dM_E} l(M_E) = \frac{d}{dM_E} \cos(M_E) \tag{71}$$

$$\frac{d}{dM_E} l(M_E) - \frac{d}{dM_E} \cos(M_E) = 0 \tag{72}$$

$$\sin(M_E) + \frac{d}{dM_E} l(M_E) = 0 \tag{73}$$

$$\int (\sin(M_E) + \frac{d}{dM_E} l(M_E)) dM_E = \int 0 dM_E \tag{74}$$

$$\int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E = \int 0 dM_E \tag{75}$$

$$y' - 1 = \int 0 dM_E - 1 \tag{76}$$

$$y' - 1 = \int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E - 1 \tag{77}$$

GPT-4

77.5 66.6 49.1 70.4

$$l(M_E) = \cos(M_E) \tag{78}$$

$$\frac{d}{dM_E} l(M_E) = \frac{d}{dM_E} \cos(M_E) \tag{79}$$

$$\frac{d}{dM_E} \cos(M_E) = -\sin(M_E) \tag{80}$$

$$\sin(M_E) + \frac{d}{dM_E} l(M_E) = \sin(M_E) - \sin(M_E) \tag{81}$$

$$\sin(M_E) + \frac{d}{dM_E} l(M_E) = 0 \tag{82}$$

$$y' - 1 = \int 0 dM_E - 1 \tag{83}$$

$$y' - 1 = \int (\sin(M_E) + \frac{d}{dM_E} \cos(M_E)) dM_E - 1 \tag{84}$$

ChatGPT

73.5 54.7 49.2 61.7

$$l(M_E) = \cos(M_E) \tag{85}$$

$$\frac{d}{dM_E} l(M_E) = -\sin(M_E) \tag{86}$$

$$\sin(M_E) + \frac{d}{dM_E} l(M_E) = 0 \tag{87}$$

$$y' - 1 = \int 0 dM_E - 1 \tag{88}$$

$$\sin(M_E)+\frac{d}{dM_E}\cos(M_E)=\sin(M_E)+\frac{d}{dM_E}l(M_E) \tag{89}$$

$$y'-1=\int(\sin(M_E)+\frac{d}{dM_E}\cos(M_E))dM_E-1 \tag{90}$$

C.1.5 Derivation 5

Correct

$$\mathrm{v_y}(L)=e^L \tag{91}$$

$$\frac{d}{dL}\mathrm{v_y}(L)=\frac{d}{dL}e^L \tag{92}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+e^L \tag{93}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=\frac{d^2}{dL^2}e^L \tag{94}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=e^L \tag{95}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+\frac{d^2}{dL^2}\mathrm{v_y}(L) \tag{96}$$

FLAN-T5

91.1 88.0 72.8 88.1

$$\mathrm{v_y}(L)=e^L \tag{97}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+e^L \tag{98}$$

$$\frac{d}{dL}2\,\mathrm{v_y}(L)=\frac{d}{dL}(\mathrm{v_y}(L)+e^L) \tag{99}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=e^L \tag{100}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=\mathrm{v_y}(L) \tag{101}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+\frac{d^2}{dL^2}\mathrm{v_y}(L) \tag{102}$$

GPT-4

91.9 87.8 68.0 88.1

$$\mathrm{v_y}(L)=e^L \tag{103}$$

$$\frac{d}{dL}\mathrm{v_y}(L)=e^L \tag{104}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=e^L \tag{105}$$

$$2\,\mathrm{v_y}(L)=2e^L \tag{106}$$

$$\mathrm{v_y}(L)+\frac{d^2}{dL^2}\mathrm{v_y}(L)=e^L+e^L \tag{107}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+\frac{d^2}{dL^2}\mathrm{v_y}(L) \tag{108}$$

ChatGPT

88.2 86.2 67.6 86.4

$$\mathrm{v_y}(L)=e^L \tag{109}$$

$$\frac{d}{dL}\mathrm{v_y}(L)=e^L \tag{110}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=e^L \tag{111}$$

$$\mathrm{v_y}(L)=\frac{d}{dL}\mathrm{v_y}(L) \tag{112}$$

$$\frac{d^2}{dL^2}\mathrm{v_y}(L)=2\frac{d}{dL}\mathrm{v_y}(L)-\mathrm{v_y}(L) \tag{113}$$

$$2\,\mathrm{v_y}(L)=\mathrm{v_y}(L)+\frac{d^2}{dL^2}\mathrm{v_y}(L) \tag{114}$$

C.1.6 Derivation 6

Correct

$$\Psi(\delta) = \log(\delta) \tag{115}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{d}{d\delta}\log(\delta) \tag{116}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{1}{\delta} \tag{117}$$

$$\frac{d}{d\delta}\log(\delta) = \frac{1}{\delta} \tag{118}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta}\frac{d}{d\delta}\log(\delta) = \frac{(\frac{d}{d\delta}\Psi(\delta))^{-\delta}}{\delta} \tag{119}$$

FLAN-T5

90.7 83.2 75.0 84.3

$$\Psi(\delta) = \log(\delta) \tag{120}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{d}{d\delta}\log(\delta) \tag{121}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{1}{\delta} \tag{122}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta}\frac{d}{d\delta}\log(\delta) = \frac{(\frac{d}{d\delta}\Psi(\delta))^{-\delta}}{\delta} \tag{123}$$

GPT-4

92.2 81.4 87.4 81.4

$$\Psi(\delta) = \log(\delta) \tag{124}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{d}{d\delta}\log(\delta) \tag{125}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{1}{\delta} \tag{126}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta}\frac{d}{d\delta}\log(\delta) = \frac{(\frac{1}{\delta})^{-\delta}}{\delta} \tag{127}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta}\frac{d}{d\delta}\log(\delta) = \frac{(\frac{d}{d\delta}\Psi(\delta))^{-\delta}}{\delta} \tag{128}$$

ChatGPT

95.0 89.4 84.2 89.4

$$\Psi(\delta) = \log(\delta) \tag{129}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{d}{d\delta}\log(\delta) \tag{130}$$

$$\frac{d}{d\delta}\Psi(\delta) = \frac{1}{\delta} \tag{131}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta} = (\frac{1}{\delta})^{-\delta} \tag{132}$$

$$(\frac{d}{d\delta}\Psi(\delta))^{-\delta}\frac{d}{d\delta}\log(\delta) = \frac{(\frac{d}{d\delta}\Psi(\delta))^{-\delta}}{\delta} \tag{133}$$

C.1.7 Derivation 7

Correct

$$\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \sin\left(P_e\right) \tag{134}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \frac{d}{dP_e}\sin\left(P_e\right) \tag{135}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \cos\left(P_e\right) \tag{136}$$

$$\frac{d}{dP_e}\sin\left(P_e\right) = \cos\left(P_e\right) \tag{137}$$

$$\frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = \frac{\cos\left(P_e\right)}{P_e} \tag{138}$$

$$-1 + \frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = -1 + \frac{\cos\left(P_e\right)}{P_e} \tag{139}$$

FLAN-T5

92.3 88.9 81.1 88.9

$$\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \sin\left(P_e\right) \tag{140}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \frac{d}{dP_e}\sin\left(P_e\right) \tag{141}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \cos\left(P_e\right) \tag{142}$$

$$\frac{\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right)}{P_e} = \frac{\cos\left(P_e\right)}{P_e} \tag{143}$$

$$\frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = \frac{\cos\left(P_e\right)}{P_e} \tag{144}$$

$$-1 + \frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = -1 + \frac{\cos\left(P_e\right)}{P_e} \tag{145}$$

GPT-4

78.4 57.7 44.0 64.3

$$\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \sin\left(P_e\right) \tag{146}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \frac{d}{dP_e}\sin\left(P_e\right) \tag{147}$$

$$\frac{d}{dP_e}\sin\left(P_e\right) = \cos\left(P_e\right) \tag{148}$$

$$-1 + \frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = -1 + \frac{\cos\left(P_e\right)}{P_e} \tag{149}$$

ChatGPT

78.4 57.7 44.0 64.3

$$\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \sin\left(P_e\right) \tag{150}$$

$$\frac{d}{dP_e}\mathrm{V}_{\mathbf{B}}\left(P_e\right) = \frac{d}{dP_e}\sin\left(P_e\right) \tag{151}$$

$$\frac{d}{dP_e}\sin\left(P_e\right) = \cos\left(P_e\right) \tag{152}$$

$$-1 + \frac{\frac{d}{dP_e}\sin\left(P_e\right)}{P_e} = -1 + \frac{\cos\left(P_e\right)}{P_e} \tag{153}$$

C.1.8 Derivation 8

Correct

$$\mathbf{M}(J) = \cos(J) \tag{154}$$

$$\int \mathbf{M}(J)dJ = \int \cos(J)dJ \tag{155}$$

$$\int \mathbf{M}(J)dJ = F_g + \sin(J) \tag{156}$$

$$F_g + \sin(J) = \int \cos(J)dJ \tag{157}$$

$$(F_g + \sin(J))^{F_g} = (\int \cos(J)dJ)^{F_g} \tag{158}$$

$$2(F_g+\sin(J))^{F_g} = (F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g} \tag{159}$$

$$\int 2(F_g+\sin(J))^{F_g}dF_g = \int((F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g}dF_g \tag{160}$$

FLAN-T5

92.1 90.9 80.3 90.9

$$\mathbf{M}(J) = \cos(J) \tag{161}$$

$$\int \mathbf{M}(J)dJ = \int \cos(J)dJ \tag{162}$$

$$(\int \mathbf{M}(J)dJ)^{F_g} = (\int \cos(J)dJ)^{F_g} \tag{163}$$

$$\int \mathbf{M}(J)dJ = F_g + \sin(J) \tag{164}$$

$$(F_g + \sin(J))^{F_g} = (\int \cos(J)dJ)^{F_g} \tag{165}$$

$$2(F_g+\sin(J))^{F_g} = (F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g} \tag{166}$$

$$\int 2(F_g+\sin(J))^{F_g}dF_g = \int((F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g}dF_g \tag{167}$$

GPT-4

86.4 73.4 71.9 76.2

$$\mathbf{M}(J) = \cos(J) \tag{168}$$

$$\int \mathbf{M}(J)dJ = \int \cos(J)dJ \tag{169}$$

$$\int \mathbf{M}(J)dJ = F_g + \sin(J) \tag{170}$$

$$2(F_g+\sin(J))^{F_g} = (F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g} \tag{171}$$

$$\int 2(F_g+\sin(J))^{F_g}dF_g = \int((F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g} \tag{172}$$

ChatGPT

84.5 74.5 67.4 76.2

$$\mathbf{M}(J) = \cos(J) \tag{173}$$

$$\int \mathbf{M}(J)dJ = \int \cos(J)dJ \tag{174}$$

$$\int \mathbf{M}(J)dJ = \sin(J) + F_g \tag{175}$$

$$F_g + \sin(J) = \int \cos(J)dJ + F_g \tag{176}$$

$$(F_g + \sin(J))^{F_g} = e^{F_g \log(F_g+\sin(J))} \tag{177}$$

$$2(F_g+\sin(J))^{F_g} = (F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g} \tag{178}$$

$$\int 2(F_g+\sin(J))^{F_g}dF_g = \int((F_g+\sin(J))^{F_g}+(\int \cos(J)dJ)^{F_g}dF_g \tag{179}$$

C.1.9 Derivation 9

Correct

$$\hat{p}_0(\phi, \mathbf{H}) = \frac{\partial}{\partial \phi}(-\mathbf{H} + \phi) \tag{179}$$

$$\hat{p}_0(\phi, \mathbf{H}) = 1 \tag{180}$$

$$\frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) = \frac{d}{d\phi}1 \tag{181}$$

$$\frac{\partial^2}{\partial \phi^2}(-\mathbf{H} + \phi) = \frac{d}{d\phi}1 \tag{182}$$

$$\frac{\partial^2}{\partial \phi^2}(-\mathbf{H} + \phi) = \frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) \tag{183}$$

$$0 = \frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) \tag{184}$$

$$0 = \frac{\partial^2}{\partial \phi^2}(-\mathbf{H} + \phi) \tag{185}$$

$$-3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)-1 = -3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial^2}{\partial \phi^2}(-\mathbf{H}+\phi)-1 \tag{186}$$

FLAN-T5

90.1 89.3 59.5 89.4

$$\hat{p}_0(\phi, \mathbf{H}) = \frac{\partial}{\partial \phi}(-\mathbf{H} + \phi) \tag{187}$$

$$\hat{p}_0(\phi, \mathbf{H}) = 1 \tag{188}$$

$$\frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) = \frac{d}{d\phi}1 \tag{189}$$

$$0 = \frac{\partial}{\partial \phi}(-\mathbf{H} + \phi) - 1 \tag{190}$$

$$-\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) = -\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial^2}{\partial \phi^2}(-\mathbf{H}+\phi) \tag{191}$$

$$0 = \frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) \tag{192}$$

$$-3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)-1 = -3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial^2}{\partial \phi^2}(-\mathbf{H}+\phi)-1 \tag{193}$$

GPT-4

70.9 43.0 32.8 53.8

$$\hat{p}_0(\phi, \mathbf{H}) = \frac{\partial}{\partial \phi}(-\mathbf{H} + \phi) \tag{194}$$

$$\hat{p}_0(\phi, \mathbf{H}) = 1 \tag{195}$$

$$0 = \frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) \tag{196}$$

$$0 = \frac{\partial}{\partial \phi}1 \tag{197}$$

$$-3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)-1 = -3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial^2}{\partial \phi^2}(-\mathbf{H}+\phi)-1 \tag{198}$$

ChatGPT

79.2 60.9 49.5 66.6

$$\hat{p}_0(\phi, \mathbf{H}) = \frac{\partial}{\partial \phi}(-\mathbf{H} + \phi) \tag{199}$$

$$\frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) = \frac{\partial^2}{\partial \phi^2}(-\mathbf{H} + \phi) \tag{200}$$

$$\hat{p}_0(\phi, \mathbf{H}) = 1 \tag{201}$$

$$0 = \frac{\partial}{\partial \phi}\hat{p}_0(\phi, \mathbf{H}) \tag{202}$$

$$-3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)-1 = -3\frac{\partial}{\partial \phi}(-\mathbf{H}+\phi)+\frac{\partial^2}{\partial \phi^2}(-\mathbf{H}+\phi)-1 \tag{203}$$

C.1.10 Derivation 10

Correct

$$\theta_1(g) = e^g \tag{204}$$

$$\int \theta_1(g)dg = \int e^g dg \tag{205}$$

$$(\int \theta_1(g)dg)^g = (\int e^g dg)^g \tag{206}$$

$$\frac{d}{dg}(\int \theta_1(g)dg)^g = \frac{d}{dg}(\int e^g dg)^g \tag{207}$$

$$\frac{d}{dg}(\int \theta_1(g)dg)^g = \frac{\partial}{\partial g}(\mathbf{g} + e^g)^g \tag{208}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + e^g)^g = \frac{d}{dg}(\int e^g dg)^g \tag{209}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g = \frac{d}{dg}(\int e^g dg)^g \tag{210}$$

FLAN-T5

90.1 80.4 77.1 82.0

$$\theta_1(g) = e^g \tag{211}$$

$$\int \theta_1(g)dg = \int e^g dg \tag{212}$$

$$(\int \theta_1(g)dg)^g = (\int e^g dg)^g \tag{213}$$

$$\frac{d}{dg}(\int \theta_1(g)dg)^g = \frac{d}{dg}(\int e^g dg)^g \tag{214}$$

$$\frac{d}{dg}(\int \theta_1(g)dg)^g = \frac{\partial}{\partial g}(\mathbf{g} + e^g)^g \tag{215}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g = \frac{d}{dg}(\int e^g dg)^g \tag{216}$$

GPT-4

75.4 79.1 63.1 79.9

$$\theta_1(g) = e^g \tag{217}$$

$$\frac{d}{dg}(\int \theta_1(g)dg)^g = \frac{d}{dg}(\int e^g dg)^g \tag{218}$$

$$\mathbf{g} + \theta_1(g) = \mathbf{g} + e^g \tag{219}$$

$$(\mathbf{g} + \theta_1(g))^g = (\mathbf{g} + e^g)^g \tag{220}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g = \frac{\partial}{\partial g}(\mathbf{g} + e^g)^g \tag{221}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g = \frac{d}{dg}(\int e^g dg)^g \tag{222}$$

ChatGPT

81.8 70.0 57.5 73.1

$$\theta_1(g) = e^g \tag{223}$$

$$\frac{d}{dg}\theta_1(g) = e^g \tag{224}$$

$$\int \theta_1(g)dg = \int e^g dg \tag{225}$$

$$(\int \theta_1(g)dg)^g = (\int e^g dg)^g \tag{226}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + e^g)^g = \frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g \tag{227}$$

$$\frac{\partial}{\partial g}(\mathbf{g} + \theta_1(g))^g = \frac{d}{dg}(\int e^g dg)^g \tag{228}$$

C.1.11 Derivation 11

Correct

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\mathbf{f}\varepsilon}{v_1} \tag{229}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\partial}{\partial \mathbf{f}} \frac{\mathbf{f}\varepsilon}{v_1} \tag{230}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\varepsilon}{v_1} \tag{231}$$

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \mathbf{f} \frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) \tag{232}$$

FLAN-T5

90.4 83.0 79.1 83.0

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\mathbf{f}\varepsilon}{v_1} \tag{233}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\partial}{\partial \mathbf{f}} \frac{\mathbf{f}\varepsilon}{v_1} \tag{234}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\varepsilon}{v_1} \tag{235}$$

$$\frac{\partial}{\partial \mathbf{f}} \frac{\mathbf{f}\varepsilon}{v_1} = \frac{\varepsilon}{v_1} \tag{236}$$

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \mathbf{f} \frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) \tag{237}$$

GPT-4

100.0 100.0 84.8 100.0

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\mathbf{f}\varepsilon}{v_1} \tag{238}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\partial}{\partial \mathbf{f}} \frac{\mathbf{f}\varepsilon}{v_1} \tag{239}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\varepsilon}{v_1} \tag{240}$$

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \mathbf{f} \frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) \tag{241}$$

ChatGPT

100.0 100.0 84.8 100.0

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\mathbf{f}\varepsilon}{v_1} \tag{242}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\partial}{\partial \mathbf{f}} \frac{\mathbf{f}\varepsilon}{v_1} \tag{243}$$

$$\frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) = \frac{\varepsilon}{v_1} \tag{244}$$

$$\dot{x}(v_1, \mathbf{f}, \varepsilon) = \mathbf{f} \frac{\partial}{\partial \mathbf{f}} \dot{x}(v_1, \mathbf{f}, \varepsilon) \tag{245}$$

C.1.12 Derivation 12

Correct

$$\omega(\mu) = e^{\mu} \tag{246}$$

$$1 = \frac{e^{\mu}}{\omega(\mu)} \tag{247}$$

$$\int 1 d\mu = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{248}$$

$$\mathbf{J} + \mu = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{249}$$

$$\mathbf{J} + \mu - \frac{1}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu - \frac{1}{\omega(\mu)} \tag{250}$$

$$\mathbf{J} + \mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} \tag{251}$$

FLAN-T5

90.0 88.4 71.1 88.4

$$\omega(\mu) = e^{\mu} \tag{252}$$

$$1 = \frac{e^{\mu}}{\omega(\mu)} \tag{253}$$

$$\int 1 d\mu = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{254}$$

$$\mathbf{J} + \mu = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{255}$$

$$\mathbf{J} + \mu = \int 1 d\mu \tag{256}$$

$$\mathbf{J} + \mu + \frac{e^{\mu}}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu + \frac{e^{\mu}}{\omega(\mu)} \tag{257}$$

$$\mathbf{J} + \mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} \tag{258}$$

GPT-4

79.5 82.3 66.9 82.9

$$\omega(\mu) = e^{\mu} \tag{259}$$

$$\frac{e^{\mu}}{\omega(\mu)} = \frac{e^{\mu}}{e^{\mu}} \tag{260}$$

$$\frac{e^{\mu}}{\omega(\mu)} = 1 \tag{261}$$

$$\int \frac{e^{\mu}}{\omega(\mu)} d\mu = \int 1 d\mu \tag{262}$$

$$\int \frac{e^{\mu}}{\omega(\mu)} d\mu = \mu + \mathbf{J} \tag{263}$$

$$\mathbf{J} + \mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} \tag{264}$$

ChatGPT

77.7 72.1 54.8 72.2

$$\omega(\mu) = e^{\mu} \tag{265}$$

$$\frac{d}{d\mu}\omega(\mu) = e^{\mu} \tag{266}$$

$$\frac{d}{d\mu} \ln \omega(\mu) = \frac{1}{\omega(\mu)} \tag{267}$$

$$\mathbf{J} + \mu = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{268}$$

$$\frac{d}{d\mu}(\mathbf{J} + \mu) = \frac{e^{\mu}}{\omega(\mu)} \tag{269}$$

$$\frac{d}{d\mu}(\mathbf{J} + \ln \omega(\mu)) = \frac{e^{\mu}}{\omega(\mu)} \tag{270}$$

$$\mathbf{J} + \ln \omega(\mu) = \int \frac{e^{\mu}}{\omega(\mu)} d\mu \tag{271}$$

$$\mathbf{J} + \mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} = \int \frac{e^{\mu}}{\omega(\mu)} d\mu + \frac{e^{\mu}}{\omega(\mu)} - \frac{1}{\omega(\mu)} \tag{272}$$

C.2 Random sample of derivations

C.2.1 Derivation 13

Correct

$$W(P_e) = \log(P_e) \tag{273}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{274}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{275}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{276}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{277}$$

FLAN-T5

99.2 99.5 66.4 99.5

$$(P_e) = \log(P_e) \tag{278}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{279}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{280}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{281}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{282}$$

GPT-4

89.3 78.9 56.8 80.7

$$W(P_e) = \log(P_e) \tag{283}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{284}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{285}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{286}$$

ChatGPT

89.3 78.9 56.8 80.7

$$W(P_e) = \log(P_e) \tag{287}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{288}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{289}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{290}$$

C.2.2 Derivation 14

Correct

$$\phi(x') = \int \log(x')dx' \tag{291}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{292}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{293}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{294}$$

$$\mathrm{t}_1(x', n_2) = \frac{d}{dx'}\phi(x') \tag{295}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{296}$$

FLAN-T5

93.0 85.7 71.8 86.5

$$\phi(x') = \int \log(x')dx' \tag{297}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{298}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{299}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{300}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{301}$$

GPT-4

90.0 82.1 60.9 82.1

$$\phi(x') = \int \log(x')dx' \tag{302}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{303}$$

$$\frac{d}{dx'}\phi(x') = \log(x') \tag{304}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{305}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{306}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\log(x')}\frac{d}{dx'}\phi(x') \tag{307}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{308}$$

ChatGPT

74.6 61.6 69.4 61.6

$$\phi(x') = \int \log(x')dx' \tag{309}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}\log(x') \tag{310}$$

$$\frac{d}{dx'}\phi(x') = \frac{1}{x'} \tag{311}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{312}$$

$$\mathrm{t}_1(x', n_2) = \log(x') + 1 - \frac{1}{x'} \tag{313}$$

$$\frac{d}{dx'}\phi(x') = \frac{1}{x'} \tag{314}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}(\log(x') + 1 - \frac{1}{x'}) \tag{315}$$

$$e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') = \frac{1}{x'}e^{-\frac{d}{dx'}\phi(x')} \tag{316}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{317}$$

C.2.3 Derivation 15

Correct

$$C(\phi_2) = \log(\phi_2) \tag{318}$$

$$2C(\phi_2) = C(\phi_2) + \log(\phi_2) \tag{319}$$

$$\frac{d}{d\phi_2} 2C(\phi_2) = \frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) \tag{320}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{321}$$

$$2\frac{d}{d\phi_2} \log(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{322}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{323}$$

FLAN-T5

99.5 99.7 79.4 99.7

$$(\phi_2) = \log(\phi_2) \tag{324}$$

$$2C(\phi_2) = C(\phi_2) + \log(\phi_2) \tag{325}$$

$$\frac{d}{d\phi_2} 2C(\phi_2) = \frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) \tag{326}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{327}$$

$$2\frac{d}{d\phi_2} \log(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{328}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{329}$$

GPT-4

85.6 79.0 68.3 79.6

$$C(\phi_2) = \log(\phi_2) \tag{330}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{1}{\phi_2} \tag{331}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = 2\frac{1}{\phi_2} \tag{332}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{333}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = 4(\frac{1}{\phi_2})^2 \tag{334}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{335}$$

ChatGPT

74.7 58.6 69.1 58.7

$$C(\phi_2) = \log(\phi_2) \tag{336}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{1}{\phi_2} \tag{337}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = 2\frac{1}{\phi_2} \tag{338}$$

$$\frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} = \frac{1}{\phi_2} + \frac{1}{\phi_2} \tag{339}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{340}$$

$$\frac{d}{d\phi_2} \log(\phi_2) = \frac{1}{\phi_2} \tag{341}$$

$$(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{1}{\phi_2})^2 \tag{342}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = 4(\frac{1}{\phi_2})^2 \tag{343}$$

$$(\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 = (\frac{1}{\phi_2} + \frac{1}{\phi_2})^2 \tag{344}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{345}$$

C.2.4 Derivation 16

Correct

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{346}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{347}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{348}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{349}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right) \tag{350}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{351}$$

FLAN-T5

87.9 74.9 80.2 77.5

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{352}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{353}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{354}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{355}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{356}$$

GPT-4

80.0 65.4 76.6 69.7

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{357}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{358}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \hat{X}^t \log\left(\hat{X}\right) \tag{359}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{360}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{361}$$

ChatGPT

88.0 82.5 76.6 83.3

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{362}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{363}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{364}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \hat{X}^t \log\left(\hat{X}\right) \tag{365}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{366}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{367}$$

C.2.5 Derivation 17

Correct

$$y(A_x) = \frac{1}{A_x} \tag{368}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{369}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log\left(A_x\right) \tag{370}$$

$$\int \frac{1}{A_x}dA_x = \varepsilon_0 + \log\left(A_x\right) \tag{371}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \varepsilon_0 + \log\left(A_x\right) - \frac{x}{A_x} \tag{372}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log\left(A_x\right) - \frac{x}{A_x}) \tag{373}$$

FLAN-T5

100.0 100.0 85.9 100.0

$$y(A_x) = \frac{1}{A_x} \tag{374}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{375}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log\left(A_x\right) \tag{376}$$

$$\int \frac{1}{A_x}dA_x = \varepsilon_0 + \log\left(A_x\right) \tag{377}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \varepsilon_0 + \log\left(A_x\right) - \frac{x}{A_x} \tag{378}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log\left(A_x\right) - \frac{x}{A_x}) \tag{379}$$

GPT-4

78.2 58.2 39.7 64.7

$$y(A_x) = \frac{1}{A_x} \tag{380}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{381}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log\left(A_x\right) \tag{382}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log\left(A_x\right) - \frac{x}{A_x}) \tag{383}$$

ChatGPT

73.1 55.5 37.2 61.7

$$y(A_x) = \frac{1}{A_x} \tag{384}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{385}$$

$$\int y(A_x)dA_x = \log\left(A_x\right) + \varepsilon_0 \tag{386}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\log\left(A_x\right) + \varepsilon_0 - \frac{x}{A_x}) \tag{387}$$

C.2.6 Derivation 18

Correct

$$u(\lambda) = \sin(\lambda) \tag{388}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{389}$$

$$\int u(\lambda) d\lambda = n - \cos(\lambda) \tag{390}$$

$$\int \sin(\lambda) d\lambda = n - \cos(\lambda) \tag{391}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{392}$$

FLAN-T5

86.0 79.2 83.3 79.2

$$u(\lambda) = \sin(\lambda) \tag{393}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{394}$$

$$\int u(\lambda) d\lambda = n - \cos(\lambda) \tag{395}$$

$$-\frac{\int u(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{396}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{397}$$

GPT-4

89.4 81.5 77.0 82.8

$$u(\lambda) = \sin(\lambda) \tag{398}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{399}$$

$$\int u(\lambda) d\lambda = n - \cos(\lambda) \tag{400}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{401}$$

ChatGPT

84.0 75.5 78.5 75.6

$$u(\lambda) = \sin(\lambda) \tag{402}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{403}$$

$$\int \sin(\lambda) d\lambda = -\cos(\lambda) + n \tag{404}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{-\cos(\lambda) + n}{\cos(\lambda)} \tag{405}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{406}$$

C.2.7 Derivation 19

Correct

$$J(\phi_1) = \sin(\phi_1) \tag{407}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{408}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{409}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{410}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{411}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{412}$$

FLAN-T5

89.4 79.0 73.5 80.8

$$(\phi_1) = \sin(\phi_1) \tag{413}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{414}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{415}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{416}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{417}$$

GPT-4

69.7 43.2 48.7 54.0

$$J(\phi_1) = \sin(\phi_1) \tag{418}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{419}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{420}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{421}$$

ChatGPT

78.8 66.9 76.2 70.7

$$J(\phi_1) = \sin(\phi_1) \tag{422}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{423}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{424}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{425}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{426}$$

C.2.8 Derivation 20

Correct

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{427}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{428}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{429}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{430}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{431}$$

FLAN-T5

80.0 65.8 59.8 70.2

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{432}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{433}$$

$$\frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{434}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{435}$$

GPT-4

82.8 66.8 65.3 71.1

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{436}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{437}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{438}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{439}$$

ChatGPT

66.2 44.0 46.8 54.4

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{440}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{441}$$

$$\sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) \tag{442}$$

C.2.9 Derivation 21

Correct

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{443}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{444}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{445}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{446}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{447}$$

FLAN-T5

100.0 99.6 76.8 99.6

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{448}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{449}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{450}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{451}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{452}$$

GPT-4

84.9 84.2 68.4 84.8

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{453}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{454}$$

$$\int e^{\Psi_\lambda} d\Psi_\lambda = e^{\Psi_\lambda} + \chi \tag{455}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + e^{\Psi_\lambda} + \chi \tag{456}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{457}$$

ChatGPT

79.8 73.6 66.7 73.6

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{458}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = \frac{d}{d\Psi_\lambda} e^{\Psi_\lambda} \tag{459}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{460}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{461}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = e^{\Psi_\lambda} + \chi \tag{462}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{463}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{464}$$

C.2.10 Derivation 22

Correct

$$\mathrm{y}'\left(\mathbf{s}\right) = \log\left(\mathbf{s}\right) \tag{465}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \int \log\left(\mathbf{s}\right) d\mathbf{s} \tag{466}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega \tag{467}$$

$$a(\mathbf{s}) = \mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{468}$$

$$a(\mathbf{s}) = -\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) \tag{469}$$

FLAN-T5

84.6 72.3 75.7 72.3

$$\mathrm{y}'\left(\mathbf{s}\right) = \log\left(\mathbf{s}\right) \tag{470}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \int \log\left(\mathbf{s}\right) d\mathbf{s} \tag{471}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega \tag{472}$$

$$\mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = -\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) \tag{473}$$

$$a(\mathbf{s}) = \mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{474}$$

$$a(\mathbf{s}) = -\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) \tag{475}$$

GPT-4

100.0 100.0 88.8 100.0

$$\mathrm{y}'\left(\mathbf{s}\right) = \log\left(\mathbf{s}\right) \tag{476}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \int \log\left(\mathbf{s}\right) d\mathbf{s} \tag{477}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega \tag{478}$$

$$a(\mathbf{s}) = \mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{479}$$

$$a(\mathbf{s}) = -\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) \tag{480}$$

ChatGPT

82.0 74.3 59.6 74.3

$$\mathrm{y}'\left(\mathbf{s}\right) = \log\left(\mathbf{s}\right) \tag{481}$$

$$\frac{d}{d\mathbf{s}}\mathrm{y}'\left(\mathbf{s}\right) = \frac{d}{d\mathbf{s}}\log\left(\mathbf{s}\right) \tag{482}$$

$$\int \frac{d}{d\mathbf{s}}\mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \int \frac{d}{d\mathbf{s}}\log\left(\mathbf{s}\right) d\mathbf{s} \tag{483}$$

$$a(\mathbf{s}) = \mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{484}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega \tag{485}$$

$$a(\mathbf{s}) = -\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) \tag{486}$$

C.2.11 Derivation 23

Correct

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{487}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{488}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{489}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{490}$$

$$\mathbf{v}(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{491}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = (\int \sin\left(F_N\right) dF_N)^2 \tag{492}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{493}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{494}$$

FLAN-T5

81.6 63.4 80.2 68.5

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{495}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{496}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{497}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{498}$$

$$\mathbf{v}(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{499}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{500}$$

GPT-4

77.6 58.5 71.0 64.7

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{501}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{502}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{503}$$

$$\int \sin\left(F_N\right) dF_N = Q - \cos\left(F_N\right) \tag{504}$$

$$\mathbf{v}(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{505}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{506}$$

ChatGPT

69.8 60.2 62.9 64.8

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{507}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{508}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{509}$$

$$\mathbf{v}(F_N) = (\frac{-\cos\left(F_N\right) + Q}{\sqrt{2}})^2 \tag{510}$$

$$\mathbf{v}(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{511}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{512}$$

C.2.12 Derivation 24

Correct

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{513}$$

$$0 = -\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right) \tag{514}$$

$$\frac{d}{d\varepsilon_0} 0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right)) \tag{515}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{516}$$

$$\int 0 d\varepsilon_0 = \int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 \tag{517}$$

FLAN-T5

100.0 99.6 85.3 99.6

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{518}$$

$$0 = -\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right) \tag{519}$$

$$\frac{d}{d\varepsilon_0} 0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right)) \tag{520}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{521}$$

$$\int 0 d\varepsilon_0 = \int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 \tag{522}$$

GPT-4

85.1 72.9 60.0 75.3

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{523}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) \tag{524}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{525}$$

$$\int 0 d\varepsilon_0 = \int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 \tag{526}$$

ChatGPT

92.8 90.0 80.5 90.1

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{527}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{528}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) \tag{529}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{530}$$

$$\int 0 d\varepsilon_0 = \int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 \tag{531}$$

C.2.13 Derivation 25

Correct

$$y(W, q, B) = W + \frac{q}{B} \tag{532}$$

$$0 = W - y(W, q, B) + \frac{q}{B} \tag{533}$$

$$\frac{d}{dq}0 = \frac{\partial}{\partial q}(W - y(W, q, B) + \frac{q}{B}) \tag{534}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{535}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{536}$$

FLAN-T5

93.4 95.2 73.3 95.3

$$y(W, q, B) = W + \frac{q}{B} \tag{537}$$

$$0 = W + \frac{q}{B} - y(W, q, B) \tag{538}$$

$$\frac{d}{dq}0 = \frac{\partial}{\partial q}(W + \frac{q}{B} - y(W, q, B)) \tag{539}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{540}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{541}$$

GPT-4

81.3 83.1 59.6 83.3

$$y(W, q, B) = W + \frac{q}{B} \tag{542}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{\partial}{\partial q}(W + \frac{q}{B}) \tag{543}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{1}{B} \tag{544}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{545}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{546}$$

ChatGPT

79.7 81.4 58.0 81.7

$$y(W, q, B) = W + \frac{q}{B} \tag{547}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{1}{B} \tag{548}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{549}$$

$$-\frac{\partial}{\partial q}(W + \frac{q}{B}) = -\frac{1}{B} \tag{550}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{551}$$

D Perturbation Analysis

The first four displayed numbers are ROUGE, BLEU, BLEURT, and GLEU differences from the static performance (positive is a decrease, negative is a score increase). The next four are the corresponding rations from Table 8.

D.1 Variable Renaming (VR)

D.1.1 Derivation 13 VR

Correct

$$\alpha(\zeta) = \log(\zeta) \tag{552}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{d}{d\zeta}\log(\zeta) \tag{553}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{1}{\zeta} \tag{554}$$

$$\frac{d}{d\zeta}\log(\zeta) = \frac{1}{\zeta} \tag{555}$$

$$\int \frac{d}{d\zeta}\log(\zeta)d\zeta = \int \frac{1}{\zeta}d\zeta \tag{556}$$

FLAN-T5

-0.8 0.2 -3.5 0.2
1.01 1.00 0.88 1.01

$$\alpha(\zeta) = \log(\zeta) \tag{557}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{d}{d\zeta}\log(\zeta) \tag{558}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{1}{\zeta} \tag{559}$$

$$\frac{d}{d\zeta}\log(\zeta) = \frac{1}{\zeta} \tag{560}$$

$$\int \frac{d}{d\zeta}\log(\zeta)d\zeta = \int \frac{1}{\zeta}d\zeta \tag{561}$$

GPT-4

-0.1 -0.1 -5.1 -0.1
0.98 0.97 1.28 0.97

$$\alpha(\zeta) = \log(\zeta) \tag{562}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{d}{d\zeta}\log(\zeta) \tag{563}$$

$$\frac{d}{d\zeta}\log(\zeta) = \frac{1}{\zeta} \tag{564}$$

$$\int \frac{d}{d\zeta}\log(\zeta)d\zeta = \int \frac{1}{\zeta}d\zeta \tag{565}$$

ChatGPT

-0.1 -0.1 -6.6 -0.1
0.98 0.97 1.28 0.97

$$\alpha(\zeta) = \log(\zeta) \tag{566}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{d}{d\zeta}\log(\zeta) \tag{567}$$

$$\frac{d}{d\zeta}\alpha(\zeta) = \frac{1}{\zeta} \tag{568}$$

$$\int \frac{d}{d\zeta}\log(\zeta)d\zeta = \int \frac{1}{\zeta}d\zeta \tag{569}$$

D.1.2 Derivation 14 VR

Correct

$$\xi(\alpha) = \int \log(\alpha) d\alpha \tag{570}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{d}{d\alpha}\int \log(\alpha) d\alpha \tag{571}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{572}$$

$$\tau(\alpha,\nu) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{573}$$

$$\tau(\alpha,\nu) = \frac{d}{d\alpha}\xi(\alpha) \tag{574}$$

$$\tau(\alpha,\nu)e^{-\frac{d}{d\alpha}\xi(\alpha)} = e^{-\frac{d}{d\alpha}\xi(\alpha)}\frac{d}{d\alpha}\xi(\alpha) \tag{575}$$

FLAN-T5

16.9 20.9 0.4 21.8
0.79 0.87 0.85 0.84

$$\xi(\alpha) = \int \log(\alpha) d\alpha \tag{576}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{d}{d\alpha}\int \log(\alpha) d\alpha \tag{577}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{578}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{579}$$

$$\tau(\alpha,\nu) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{580}$$

$$\tau(\alpha,\nu) = \frac{d}{d\alpha}\int \log(\alpha) d\alpha \tag{581}$$

$$\tau(\alpha,\nu)e^{-\frac{d}{d\alpha}\int \log(\alpha) d\alpha} = e^{-\frac{d}{d\alpha}\int \log(\alpha) d\alpha}\frac{d}{d\alpha}\int \log(\alpha) d\alpha \tag{582}$$

GPT-4

-5.8 -14.9 -16.3 -14.9
0.93 0.77 -7.62 0.85

$$\xi(\alpha) = \int \log(\alpha) d\alpha \tag{583}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{d}{d\alpha}\int \log(\alpha) d\alpha \tag{584}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \log(\alpha) \tag{585}$$

$$\tau(\alpha,\nu) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{586}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{587}$$

$$\tau(\alpha,\nu)e^{-\frac{d}{d\alpha}\xi(\alpha)} = e^{-\frac{d}{d\alpha}\xi(\alpha)}\frac{d}{d\alpha}\xi(\alpha) \tag{588}$$

ChatGPT

-12.2 -19.6 14.4 -19.6
0.82 0.69 0.21 0.80

$$\xi(\alpha) = \int \log(\alpha) d\alpha \tag{589}$$

$$\frac{d}{d\alpha}\xi(\alpha) = \log(\alpha) \tag{590}$$

$$\tau(\alpha,\nu) = \frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{591}$$

$$\tau(\alpha,\nu) = \log(\alpha) + 1 \tag{592}$$

$$e^{-\frac{d}{d\alpha}\xi(\alpha)} = \frac{1}{\alpha} \tag{593}$$

$$\tau(\alpha,\nu)e^{-\frac{d}{d\alpha}\xi(\alpha)} = e^{-\frac{d}{d\alpha}\xi(\alpha)}\frac{\partial}{\partial\alpha}(\alpha\log(\alpha) - \alpha + \nu) \tag{594}$$

$$\int \log(\alpha) d\alpha \tau(\alpha,\nu)e^{-\frac{d}{d\alpha}\xi(\alpha)} = e^{-\frac{d}{d\alpha}\xi(\alpha)}\frac{d}{d\alpha}\xi(\alpha) \tag{595}$$

D.1.3 Derivation 15 VR

Correct

$$\alpha(\iota) = \log(\iota) \tag{596}$$

$$2\alpha(\iota) = \alpha(\iota) + \log(\iota) \tag{597}$$

$$\frac{d}{d\iota}2\alpha(\iota) = \frac{d}{d\iota}(\alpha(\iota) + \log(\iota)) \tag{598}$$

$$2\frac{d}{d\iota}\alpha(\iota) = \frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} \tag{599}$$

$$2\frac{d}{d\iota}\log(\iota) = \frac{d}{d\iota}\log(\iota) + \frac{1}{\iota} \tag{600}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = (\frac{d}{d\iota}\log(\iota) + \frac{1}{\iota})^2 \tag{601}$$

FLAN-T5

-0.5 0.1 -4.9 0.1
1.01 1.00 0.81 1.00

$$\alpha(\iota) = \log(\iota) \tag{602}$$

$$2\alpha(\iota) = \alpha(\iota) + \log(\iota) \tag{603}$$

$$\frac{d}{d\iota}2\alpha(\iota) = \frac{d}{d\iota}(\alpha(\iota) + \log(\iota)) \tag{604}$$

$$2\frac{d}{d\iota}\alpha(\iota) = \frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} \tag{605}$$

$$2\frac{d}{d\iota}\log(\iota) = \frac{d}{d\iota}\log(\iota) + \frac{1}{\iota} \tag{606}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = (\frac{d}{d\iota}\log(\iota) + \frac{1}{\iota})^2 \tag{607}$$

GPT-4

8.6 10.1 1.1 10.6
1.17 0.89 1.06 0.87

$$\alpha(\iota) = \log(\iota) \tag{608}$$

$$\frac{d}{d\iota}\alpha(\iota) = \frac{1}{\iota} \tag{609}$$

$$2\frac{d}{d\iota}\alpha(\iota) = 2\frac{1}{\iota} \tag{610}$$

$$\frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} = \frac{1}{\iota} + \frac{1}{\iota} \tag{611}$$

$$2\frac{d}{d\iota}\alpha(\iota) = \frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} \tag{612}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = 4(\frac{1}{\iota})^2 \tag{613}$$

$$(\frac{d}{d\iota}\log(\iota) + \frac{1}{\iota})^2 = (2\frac{1}{\iota})^2 \tag{614}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = (\frac{d}{d\iota}\log(\iota) + \frac{1}{\iota})^2 \tag{615}$$

ChatGPT

-2.0 -6.8 -6.2 -6.8
1.02 0.86 0.35 0.88

$$\alpha(\iota) = \log(\iota) \tag{616}$$

$$\frac{d}{d\iota}\alpha(\iota) = \frac{d}{d\iota}\log(\iota) \tag{617}$$

$$2\frac{d}{d\iota}\alpha(\iota) = 2\frac{d}{d\iota}\log(\iota) \tag{618}$$

$$\frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} = \frac{d}{d\iota}\log(\iota) + \frac{1}{\iota} \tag{619}$$

$$2\frac{d}{d\iota}\alpha(\iota) = \frac{d}{d\iota}\alpha(\iota) + \frac{1}{\iota} \tag{620}$$

$$(\frac{d}{d\iota}\log(\iota))^2 = (\frac{d}{d\iota}\log(\iota))^2 \tag{621}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = 4(\frac{d}{d\iota}\log(\iota))^2 \tag{622}$$

$$4(\frac{d}{d\iota}\log(\iota))^2 = (\frac{d}{d\iota}\log(\iota) + \frac{1}{\iota})^2 \tag{623}$$

D.1.4 Derivation 16 VR

Correct

$$\upsilon(\xi, \kappa) = \xi^{\kappa} \tag{624}$$

$$\frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{625}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{626}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{627}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \upsilon(\xi, \kappa)\log\left(\xi\right) \tag{628}$$

$$\xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{629}$$

FLAN-T5

3.6 0.9 17.1 2.1
1.07 1.00 1.71 0.99

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{630}$$

$$\frac{\partial}{\partial \kappa}(xi + \frac{\partial}{\partial \kappa}\xi^{\kappa}) = \frac{\partial}{\partial \kappa}(xi + \frac{\partial}{\partial \kappa}\xi^{\kappa}) \tag{631}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{632}$$

$$\xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{633}$$

GPT-4

-5.1 -5.6 6.8 -4.2
1.09 1.04 1.50 1.02

$$\upsilon(\xi, \kappa) = \xi^{\kappa} \tag{634}$$

$$\frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{635}$$

$$\frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi^{\kappa}\log\left(\xi\right) \tag{636}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{637}$$

$$\xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{638}$$

ChatGPT

4.7 13.3 2.9 10.9
0.97 0.83 1.89 0.90

$$\upsilon(\xi, \kappa) = \xi^{\kappa} \tag{639}$$

$$\frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \frac{\partial}{\partial \kappa}\xi^{\kappa} \tag{640}$$

$$\frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi^{\kappa}\log\left(\xi\right) \tag{641}$$

$$\xi + \frac{\partial}{\partial \kappa}\upsilon(\xi, \kappa) = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{642}$$

$$\xi + \frac{\partial}{\partial \kappa}\xi^{\kappa} = \xi + \xi^{\kappa}\log\left(\xi\right) \tag{643}$$

D.1.5 Derivation 17 VR

Correct

$$\gamma(\zeta) = \frac{1}{\zeta} \tag{644}$$

$$\int \gamma(\zeta) d\zeta = \int \frac{1}{\zeta} d\zeta \tag{645}$$

$$\int \gamma(\zeta) d\zeta = \upsilon + \log(\zeta) \tag{646}$$

$$\int \frac{1}{\zeta} d\zeta = \upsilon + \log(\zeta) \tag{647}$$

$$-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta = -\frac{\beta}{\zeta} + \upsilon + \log(\zeta) \tag{648}$$

$$\frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta) = \frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \upsilon + \log(\zeta)) \tag{649}$$

FLAN-T5

0.0 0.4 8.8 0.4
1.00 1.00 0.60 1.00

$$\gamma(\zeta) = \frac{1}{\zeta} \tag{650}$$

$$\int \gamma(\zeta) d\zeta = \int \frac{1}{\zeta} d\zeta \tag{651}$$

$$\int \gamma(\zeta) d\zeta = \upsilon + \log(\zeta) \tag{652}$$

$$\int \frac{1}{\zeta} d\zeta = \upsilon + \log(\zeta) \tag{653}$$

$$-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta = -\frac{\beta}{\zeta} + \upsilon + \log(\zeta) \tag{654}$$

$$\frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta) = \frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \upsilon + \log(\zeta)) \tag{655}$$

GPT-4

-0.8 -1.6 -4.6 -1.1
1.09 1.04 1.69 1.04

$$\gamma(\zeta) = \frac{1}{\zeta} \tag{656}$$

$$\int \gamma(\zeta) d\zeta = \int \frac{1}{\zeta} d\zeta \tag{657}$$

$$\int \gamma(\zeta) d\zeta = \upsilon + \log(\zeta) \tag{658}$$

$$\frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta) = \frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \upsilon + \log(\zeta)) \tag{659}$$

ChatGPT

-12.0 -28.0 -30.2 -22.3
0.94 0.98 0.05 0.98

$$\gamma(\zeta) = \frac{1}{\zeta} \tag{660}$$

$$\int \gamma(\zeta) d\zeta = \log(\zeta) + \upsilon \tag{661}$$

$$\int \gamma(\zeta) d\zeta = \int \frac{1}{\zeta} d\zeta + \upsilon \tag{662}$$

$$-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta = -\frac{\beta}{\zeta} + \log(\zeta) + \upsilon \tag{663}$$

$$\frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \int \frac{1}{\zeta} d\zeta) = \frac{\partial}{\partial \beta}(-\frac{\beta}{\zeta} + \log(\zeta) + \upsilon) \tag{664}$$

D.1.6 Derivation 18 VR

Correct

$$\tau(\kappa) = \sin(\kappa) \tag{665}$$

$$\int \tau(\kappa) d\kappa = \int \sin(\kappa) d\kappa \tag{666}$$

$$\int \tau(\kappa) d\kappa = \zeta - \cos(\kappa) \tag{667}$$

$$\int \sin(\kappa) d\kappa = \zeta - \cos(\kappa) \tag{668}$$

$$-\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = -\frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{669}$$

FLAN-T5

7.7 15.5 -4.6 15.6

0.78 0.76 1.25 0.77

$$\tau(\kappa) = \sin(\kappa) \tag{670}$$

$$\int \tau(\kappa) d\kappa = \int \sin(\kappa) d\kappa \tag{671}$$

$$\int \tau(\kappa) d\kappa = \zeta - \cos(\kappa) \tag{672}$$

$$\frac{\int \tau(\kappa) d\kappa}{\cos(\kappa)} = \frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{673}$$

$$\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = \frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{674}$$

$$-\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = -\frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{675}$$

GPT-4

-10.6 -18.5 -8.5 -17.2

0.83 0.83 1.21 0.83

$$\tau(\kappa) = \sin(\kappa) \tag{676}$$

$$\int \tau(\kappa) d\kappa = \int \sin(\kappa) d\kappa \tag{677}$$

$$\int \tau(\kappa) d\kappa = \zeta - \cos(\kappa) \tag{678}$$

$$\int \sin(\kappa) d\kappa = \zeta - \cos(\kappa) \tag{679}$$

$$-\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = -\frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{680}$$

ChatGPT

0.0 -0.5 -3.4 -0.5

1.22 1.10 1.16 1.09

$$\tau(\kappa) = \sin(\kappa) \tag{681}$$

$$\int \tau(\kappa) d\kappa = \int \sin(\kappa) d\kappa \tag{682}$$

$$\int \sin(\kappa) d\kappa = -\cos(\kappa) + \zeta \tag{683}$$

$$-\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = -\frac{-\cos(\kappa) + \zeta}{\cos(\kappa)} \tag{684}$$

$$-\frac{\int \sin(\kappa) d\kappa}{\cos(\kappa)} = -\frac{\zeta - \cos(\kappa)}{\cos(\kappa)} \tag{685}$$

D.1.7 Derivation 19 VR

Correct

$$\gamma(\xi) = \sin(\xi) \tag{686}$$

$$\frac{d}{d\xi}\gamma(\xi) = \frac{d}{d\xi}\sin(\xi) \tag{687}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\frac{d}{d\xi}\sin(\xi) \tag{688}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\cos(\xi) \tag{689}$$

$$\sin(\xi)\frac{d}{d\xi}\sin(\xi) = \sin(\xi)\cos(\xi) \tag{690}$$

$$\gamma(\xi)\frac{d}{d\xi}\gamma(\xi) = \gamma(\xi)\cos(\xi) \tag{691}$$

FLAN-T5

-0.7 -0.5 -5.1 -0.4
1.03 0.98 0.69 0.98

$$\gamma(\xi) = \sin(\xi) \tag{692}$$

$$\frac{d}{d\xi}\gamma(\xi) = \frac{d}{d\xi}\sin(\xi) \tag{693}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\frac{d}{d\xi}\sin(\xi) \tag{694}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\cos(\xi) \tag{695}$$

$$\gamma(\xi)\frac{d}{d\xi}\gamma(\xi) = \gamma(\xi)\cos(\xi) \tag{696}$$

GPT-4

-11.6 -22.6 -25.2 -16.1
0.73 0.86 0.64 0.85

$$\gamma(\xi) = \sin(\xi) \tag{697}$$

$$\frac{d}{d\xi}\gamma(\xi) = \frac{d}{d\xi}\sin(\xi) \tag{698}$$

$$\frac{d}{d\xi}\sin(\xi) = \cos(\xi) \tag{699}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\cos(\xi) \tag{700}$$

$$\gamma(\xi)\frac{d}{d\xi}\gamma(\xi) = \gamma(\xi)\cos(\xi) \tag{701}$$

ChatGPT

2.4 -0.8 6.9 -0.5
0.78 0.92 0.48 0.93

$$\gamma(\xi) = \sin(\xi) \tag{702}$$

$$\frac{d}{d\xi}\gamma(\xi) = \cos(\xi) \tag{703}$$

$$\sin(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\cos(\xi) \tag{704}$$

$$\gamma(\xi)\frac{d}{d\xi}\gamma(\xi) = \sin(\xi)\cos(\xi) \tag{705}$$

$$\gamma(\xi)\frac{d}{d\xi}\gamma(\xi) = \gamma(\xi)\cos(\xi) \tag{706}$$

D.1.8 Derivation 20 VR

Correct

$$v(\xi) = \frac{d}{d\xi} \sin(\xi) \tag{707}$$

$$v(\xi) = \cos(\xi) \tag{708}$$

$$v(\xi) \sin(\xi) = \sin(\xi) \frac{d}{d\xi} \sin(\xi) \tag{709}$$

$$\cos(\xi) = \frac{d}{d\xi} \sin(\xi) \tag{710}$$

$$v(\xi) \sin(\xi) = \sin(\xi) \cos(\xi) \tag{711}$$

FLAN-T5

3.8 -4.7 -10.2 -0.4
0.75 1.19 0.41 1.01

$$\xi + v(\xi) = \xi + \frac{d}{d\xi} \sin(\xi) \tag{712}$$

$$\frac{d}{d\xi}(\xi + v(\xi)) = \frac{d}{d\xi}(\xi + \frac{d}{d\xi} \sin(\xi)) \tag{713}$$

$$v(\xi) = \cos(\xi) \tag{714}$$

$$\cos(\xi) = \frac{d}{d\xi} \sin(\xi) \tag{715}$$

$$\frac{d}{d\xi} \sin(\xi) = \sin(\xi) \cos(\xi) \tag{716}$$

$$v(\xi) \sin(\xi) = \sin(\xi) \cos(\xi) \tag{717}$$

GPT-4

17.1 26.6 18.4 19.3
0.45 0.50 1.02 0.70

$$v(\xi) = \frac{d}{d\xi} \sin(\xi) \tag{718}$$

$$v(\xi) = \cos(\xi) \tag{719}$$

$$v(\xi) \sin(\xi) = \sin(\xi) \cos(\xi) \tag{720}$$

ChatGPT

-17.0 -34.7 -32.0 -25.9
0.69 1.13 0.64 0.97

$$v(\xi) = \frac{d}{d\xi} \sin(\xi) \tag{721}$$

$$v(\xi) = \cos(\xi) \tag{722}$$

$$\sin(\xi) \cos(\xi) = \frac{d}{d\xi} \sin(\xi) \sin(\xi) \tag{723}$$

$$v(\xi) \sin(\xi) = \sin(\xi) \cos(\xi) \tag{724}$$

D.1.9 Derivation 21 VR

Correct

$$\nu(\tau) = e^{\tau} \tag{725}$$

$$\int \nu(\tau) d\tau = \int e^{\tau} d\tau \tag{726}$$

$$\tau + \int \nu(\tau) d\tau = \tau + \int e^{\tau} d\tau \tag{727}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + e^{\tau} \tag{728}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + \nu(\tau) \tag{729}$$

FLAN-T5

27.1 36.7 3.4 36.7
0.80 1.34 0.64 1.14

$$\nu(\tau) = e^{\tau} \tag{730}$$

$$\int \nu(\tau) d\tau = \int e^{\tau} d\tau \tag{731}$$

$$\frac{d}{d\tau} \int \nu(\tau) d\tau = \frac{d}{d\tau} \int e^{\tau} d\tau \tag{732}$$

$$\int \frac{d}{d\tau} \int \nu(\tau) d\tau d\tau = \int \frac{d}{d\tau} \int e^{\tau} d\tau d\tau \tag{733}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + e^{\tau} \tag{734}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + \nu(\tau) \tag{735}$$

GPT-4

-0.2 -2.6 -11.3 -2.3
1.11 1.03 0.77 1.01

$$\nu(\tau) = e^{\tau} \tag{736}$$

$$\int \nu(\tau) d\tau = \int e^{\tau} d\tau \tag{737}$$

$$\int \nu(\tau) d\tau = e^{\tau} + \gamma \tag{738}$$

$$\tau + \int \nu(\tau) d\tau = \tau + e^{\tau} + \gamma \tag{739}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + \nu(\tau) \tag{740}$$

ChatGPT

0.6 -11.0 0.9 -11.2
1.14 0.53 1.20 0.75

$$\frac{d}{d\tau} \nu(\tau) = e^{\tau} \tag{741}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + e^{\tau} \tag{742}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + \int \frac{d}{d\tau} \nu(\tau) d\tau \tag{743}$$

$$\tau + \int \nu(\tau) d\tau = \gamma + \tau + \nu(\tau) \tag{744}$$

D.1.10 Derivation 22 VR

Correct

$$\nu(\xi) = \log(\xi) \tag{745}$$

$$\int \nu(\xi) d\xi = \int \log(\xi) d\xi \tag{746}$$

$$\int \nu(\xi) d\xi = \kappa + \xi \log(\xi) - \xi \tag{747}$$

$$\tau(\xi) = \nu(\xi) - \int \nu(\xi) d\xi \tag{748}$$

$$\tau(\xi) = -\kappa - \xi \log(\xi) + \xi + \nu(\xi) \tag{749}$$

FLAN-T5

14.3 15.9 1.1 15.9
0.65 1.09 0.55 0.99

$$\nu(\xi) = \log(\xi) \tag{750}$$

$$\int \nu(\xi) d\xi = \int \log(\xi) d\xi \tag{751}$$

$$\int \nu(\xi) d\xi = \kappa + \xi \log(\xi) - \xi \tag{752}$$

$$\int \log(\xi) d\xi = \kappa + \xi \log(\xi) - \xi \tag{753}$$

$$\xi + \int \log(\xi) d\xi = \kappa + \xi \log(\xi) - \xi + \int \log(\xi) d\xi \tag{754}$$

$$\mu(\xi) = \nu(\xi) - \int \nu(\xi) d\xi \tag{755}$$

$$\tau(\xi) = \nu(\xi) - \int \log(\xi) d\xi \tag{756}$$

$$\tau(\xi) = -\kappa - \xi \log(\xi) + \xi + \nu(\xi) \tag{757}$$

GPT-4

0.0 0.0 -3.3 0.0
1.00 1.00 1.00 1.00

$$\nu(\xi) = \log(\xi) \tag{758}$$

$$\int \nu(\xi) d\xi = \int \log(\xi) d\xi \tag{759}$$

$$\int \nu(\xi) d\xi = \kappa + \xi \log(\xi) - \xi \tag{760}$$

$$\tau(\xi) = \nu(\xi) - \int \nu(\xi) d\xi \tag{761}$$

$$\tau(\xi) = -\kappa - \xi \log(\xi) + \xi + \nu(\xi) \tag{762}$$

ChatGPT

15.3 19.1 -11.7 19.0
2.21 1.35 0.45 1.19

$$\nu(\xi) = \log(\xi) \tag{763}$$

$$\frac{d}{d\xi} \nu(\xi) = \frac{1}{\xi} \tag{764}$$

$$\tau(\xi) = \nu(\xi) - \int \nu(\xi) d\xi \tag{765}$$

$$\frac{d}{d\xi} \tau(\xi) = \frac{d}{d\xi} \nu(\xi) - \frac{d}{d\xi} \int \nu(\xi) d\xi \tag{766}$$

$$\frac{d}{d\xi} \tau(\xi) = \frac{1}{\xi} - \nu(\xi) \tag{767}$$

$$\int \nu(\xi) d\xi = \kappa + \xi \log(\xi) - \xi \tag{768}$$

$$\tau(\xi) = -\kappa - \xi \log(\xi) + \xi + \nu(\xi) \tag{769}$$

D.1.11 Derivation 23 VR

Correct

$$\iota(\alpha) = \sin(\alpha) \tag{770}$$

$$\int \iota(\alpha) d\alpha = \int \sin(\alpha) d\alpha \tag{771}$$

$$\upsilon(\alpha) = (\int \iota(\alpha) d\alpha)^2 \tag{772}$$

$$\upsilon(\alpha) = (\int \sin(\alpha) d\alpha)^2 \tag{773}$$

$$\upsilon(\alpha) = (\xi - \cos(\alpha))^2 \tag{774}$$

$$(\int \iota(\alpha) d\alpha)^2 = (\int \sin(\alpha) d\alpha)^2 \tag{775}$$

$$(\int \iota(\alpha) d\alpha)^2 = (\xi - \cos(\alpha))^2 \tag{776}$$

$$(\int \sin(\alpha) d\alpha)^2 = (\xi - \cos(\alpha))^2 \tag{777}$$

FLAN-T5

-0.3 -0.3 14.5 0.2
0.82 0.94 2.55 0.94

$$\iota(\alpha) = \sin(\alpha) \tag{778}$$

$$\int \iota(\alpha) d\alpha = \int \sin(\alpha) d\alpha \tag{779}$$

$$(\int \iota(\alpha) d\alpha)^2 = (\int \sin(\alpha) d\alpha)^2 \tag{780}$$

$$\upsilon(\alpha) = (\int \iota(\alpha) d\alpha)^2 \tag{781}$$

$$\upsilon(\alpha) = (\xi - \cos(\alpha))^2 \tag{782}$$

$$\int \sin(\alpha) d\alpha = (\xi - \cos(\alpha))^2 \tag{783}$$

GPT-4

0.6 -2.5 6.2 -1.7
0.88 1.04 1.65 1.03

$$\iota(\alpha) = \sin(\alpha) \tag{784}$$

$$\upsilon(\alpha) = (\int \iota(\alpha) d\alpha)^2 \tag{785}$$

$$\upsilon(\alpha) = (\int \sin(\alpha) d\alpha)^2 \tag{786}$$

$$\int \sin(\alpha) d\alpha = \xi - \cos(\alpha) \tag{787}$$

$$\upsilon(\alpha) = (\xi - \cos(\alpha))^2 \tag{788}$$

$$(\int \sin(\alpha) d\alpha)^2 = (\xi - \cos(\alpha))^2 \tag{789}$$

ChatGPT

-9.9 -23.2 -13.3 -19.0
0.67 1.11 1.53 1.05

$$\iota(\alpha) = \sin(\alpha) \tag{790}$$

$$\upsilon(\alpha) = (\int \iota(\alpha) d\alpha)^2 \tag{791}$$

$$\upsilon(\alpha) = (\int \sin(\alpha) d\alpha)^2 \tag{792}$$

$$\upsilon(\alpha) = (\xi - \cos(\alpha))^2 \tag{793}$$

$$\int \iota(\alpha) d\alpha = -\cos(\alpha) + \xi \tag{794}$$

$$\upsilon(\alpha) = (\int \iota(\alpha) d\alpha)^2 \tag{795}$$

$$\upsilon(\alpha) = (\xi - \cos(\alpha))^2 \tag{796}$$

$$(\int \sin(\alpha) d\alpha)^2 = (\xi - \cos(\alpha))^2 \tag{797}$$

D.1.12 Derivation 24 VR

Correct

$$\alpha(\upsilon) = \sin(\upsilon) \tag{798}$$

$$0 = -\alpha(\upsilon) + \sin(\upsilon) \tag{799}$$

$$\frac{d}{d\upsilon}0 = \frac{d}{d\upsilon}(-\alpha(\upsilon) + \sin(\upsilon)) \tag{800}$$

$$0 = \cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon) \tag{801}$$

$$\int 0 d\upsilon = \int (\cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon)) d\upsilon \tag{802}$$

FLAN-T5

0.0 0.2 9.3 0.2
1.00 1.00 1.52 1.00

$$\alpha(\upsilon) = \sin(\upsilon) \tag{803}$$

$$0 = -\alpha(\upsilon) + \sin(\upsilon) \tag{804}$$

$$\frac{d}{d\upsilon}0 = \frac{d}{d\upsilon}(-\alpha(\upsilon) + \sin(\upsilon)) \tag{805}$$

$$0 = \cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon) \tag{806}$$

$$\int 0 d\upsilon = \int (\cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon)) d\upsilon \tag{807}$$

GPT-4

-0.9 -13.5 -9.8 -11.2
0.86 1.08 0.73 1.02

$$\alpha(\upsilon) = \sin(\upsilon) \tag{808}$$

$$\frac{d}{d\upsilon}\alpha(\upsilon) = \frac{d}{d\upsilon}\sin(\upsilon) \tag{809}$$

$$\frac{d}{d\upsilon}\alpha(\upsilon) = \cos(\upsilon) \tag{810}$$

$$0 = \cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon) \tag{811}$$

$$\int 0 d\upsilon = \int (\cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon)) d\upsilon \tag{812}$$

ChatGPT

10.1 19.9 25.8 17.3
0.82 0.60 4.29 0.73

$$\alpha(\upsilon) = \sin(\upsilon) \tag{813}$$

$$\frac{d}{d\upsilon}\alpha(\upsilon) = \cos(\upsilon) \tag{814}$$

$$0 = \cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon) \tag{815}$$

$$\int 0 d\upsilon = \int (\cos(\upsilon) - \frac{d}{d\upsilon}\alpha(\upsilon)) d\upsilon \tag{816}$$

D.1.13 Derivation 25 VR

Correct

$$\upsilon(\kappa, \tau, \upsilon) = \frac{\kappa}{\upsilon} + \tau \tag{817}$$

$$0 = \frac{\kappa}{\upsilon} + \tau - \upsilon(\kappa, \tau, \upsilon) \tag{818}$$

$$\frac{d}{d\kappa}0 = \frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau - \upsilon(\kappa, \tau, \upsilon)) \tag{819}$$

$$0 = -\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) + \frac{1}{\upsilon} \tag{820}$$

$$0 = -\frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau) + \frac{1}{\upsilon} \tag{821}$$

FLAN-T5

19.8 24.7 -3.6 24.5
0.92 0.82 0.57 0.81

$$\frac{\kappa}{\upsilon} + \tau \tag{822}$$

$$0 = -\frac{\kappa}{\upsilon} + \tau - \frac{1}{\upsilon} \tag{823}$$

$$\frac{d}{d\kappa}0 = \frac{\partial}{\partial\kappa}(-\frac{\kappa}{\upsilon} + \tau - \frac{1}{\upsilon}) \tag{824}$$

$$\frac{d}{d\kappa}0 = \frac{\partial}{\partial\kappa}(-\frac{\kappa}{\upsilon} + \tau - \frac{1}{\upsilon}) \tag{825}$$

$$0 = -\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) + \frac{1}{\upsilon} \tag{826}$$

$$0 = -\frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau) + \frac{1}{\upsilon} \tag{827}$$

GPT-4

3.3 -2.5 -14.1 -2.5
1.10 1.12 0.55 1.09

$$\upsilon(\kappa, \tau, \upsilon) = \frac{\kappa}{\upsilon} + \tau \tag{828}$$

$$\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) = \frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau) \tag{829}$$

$$\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) = \frac{1}{\upsilon} \tag{830}$$

$$0 = -\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) + \frac{1}{\upsilon} \tag{831}$$

$$0 = -\frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau) + \frac{1}{\upsilon} \tag{832}$$

ChatGPT

3.4 -1.5 -14.2 -1.7
1.13 1.13 0.57 1.10

$$\upsilon(\kappa, \tau, \upsilon) = \frac{\kappa}{\upsilon} + \tau \tag{833}$$

$$\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) = \frac{1}{\upsilon} \tag{834}$$

$$-\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) = -\frac{1}{\upsilon} \tag{835}$$

$$0 = -\frac{\partial}{\partial\kappa}\upsilon(\kappa, \tau, \upsilon) + \frac{1}{\upsilon} \tag{836}$$

$$0 = -\frac{\partial}{\partial\kappa}(\frac{\kappa}{\upsilon} + \tau) + \frac{1}{\upsilon} \tag{837}$$

D.2 Expression Exchange (EE)

D.2.1 Derivation 13 EE

Correct

$$\log\left(P_e\right) = W(P_e) \tag{838}$$

$$\frac{d}{dP_e}\log\left(P_e\right) = \frac{d}{dP_e}W(P_e) \tag{839}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}W(P_e) \tag{840}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}\log\left(P_e\right) \tag{841}$$

$$\int \frac{1}{P_e}dP_e = \int \frac{d}{dP_e}\log\left(P_e\right)dP_e \tag{842}$$

FLAN-T5

-0.8 0.0 0.5 0.0
0.99 0.99 0.96 0.99

$$\log\left(P_e\right) = W(P_e) \tag{843}$$

$$\frac{d}{dP_e}\log\left(P_e\right) = \frac{d}{dP_e}W(P_e) \tag{844}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}W(P_e) \tag{845}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}\log\left(P_e\right) \tag{846}$$

$$\int \frac{1}{P_e}dP_e = \int \frac{d}{dP_e}\log\left(P_e\right)dP_e \tag{847}$$

GPT-4

-0.7 -11.9 -18.7 -10.1
0.85 0.76 1.15 0.76

$$\log\left(P_e\right) = W(P_e) \tag{848}$$

$$\frac{d}{dP_e}\log\left(P_e\right) = \frac{d}{dP_e}W(P_e) \tag{849}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}W(P_e) \tag{850}$$

$$\int \frac{1}{P_e}dP_e = \int \frac{d}{dP_e}W(P_e)dP_e \tag{851}$$

$$\int \frac{1}{P_e}dP_e = \int \frac{d}{dP_e}\log\left(P_e\right)dP_e \tag{852}$$

ChatGPT

-10.7 -20.4 -15.6 -18.6
0.91 0.82 1.14 0.82

$$\log\left(P_e\right) = W(P_e) \tag{853}$$

$$\frac{d}{dP_e}\log\left(P_e\right) = \frac{1}{P_e} \tag{854}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log\left(P_e\right) \tag{855}$$

$$\frac{1}{P_e} = \frac{d}{dP_e}W(P_e) \tag{856}$$

$$\int \frac{1}{P_e}dP_e = \int \frac{d}{dP_e}\log\left(P_e\right)dP_e \tag{857}$$

D.2.2 Derivation 14 EE

Correct

$$\int \log\left(x'\right) dx' = \phi(x') \tag{858}$$

$$\frac{d}{dx'} \int \log\left(x'\right) dx' = \frac{d}{dx'} \phi(x') \tag{859}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \frac{d}{dx'} \phi(x') \tag{860}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \mathrm{t}_1\left(x', n_2\right) \tag{861}$$

$$\frac{d}{dx'} \phi(x') = \mathrm{t}_1\left(x', n_2\right) \tag{862}$$

$$e^{-\frac{d}{dx'}\phi(x')} \frac{d}{dx'} \phi(x') = \mathrm{t}_1\left(x', n_2\right) e^{-\frac{d}{dx'}\phi(x')} \tag{863}$$

FLAN-T5

-0.2 -6.0 13.0 -5.1
0.93 0.82 0.57 0.82

$$\int \log\left(x'\right) dx' = \phi(x') \tag{864}$$

$$\frac{d}{dx'} \int \log\left(x'\right) dx' = \frac{d}{dx'} \phi(x') \tag{865}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \frac{d}{dx'} \phi(x') \tag{866}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \frac{d}{dx'} \phi(x') \tag{867}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \mathrm{t}_1\left(x', n_2\right) \tag{868}$$

$$e^{-\frac{d}{dx'}\phi(x')} \frac{d}{dx'} \phi(x') = \mathrm{t}_1\left(x', n_2\right) e^{-\frac{d}{dx'}\phi(x')} \tag{869}$$

GPT-4

7.2 13.9 5.3 9.9
0.76 0.52 0.64 0.60

$$\int \log\left(x'\right) dx' = \phi(x') \tag{870}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \mathrm{t}_1\left(x', n_2\right) \tag{871}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \frac{d}{dx'} \phi(x') \tag{872}$$

$$e^{-\frac{d}{dx'}\phi(x')} \frac{d}{dx'} \phi(x') = \mathrm{t}_1\left(x', n_2\right) e^{-\frac{d}{dx'}\phi(x')} \tag{873}$$

ChatGPT

-13.5 -23.9 1.5 -23.9
0.79 0.63 0.83 0.67

$$\int \log\left(x'\right) dx' = \phi(x') \tag{874}$$

$$\frac{d}{dx'} \int \log\left(x'\right) dx' = \frac{d}{dx'} \phi(x') \tag{875}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \mathrm{t}_1\left(x', n_2\right) \tag{876}$$

$$\frac{\partial}{\partial x'}(n_2 + x' \log\left(x'\right) - x') = \frac{d}{dx'}(x' \log\left(x'\right)) - 1 \tag{877}$$

$$\frac{d}{dx'}(x' \log\left(x'\right)) = \log\left(x'\right) + 1 \tag{878}$$

$$\frac{d}{dx'} \phi(x') = \frac{1}{x'} \tag{879}$$

$$e^{-\frac{d}{dx'}\phi(x')} \frac{d}{dx'} \phi(x') = \mathrm{t}_1\left(x', n_2\right) e^{-\frac{d}{dx'}\phi(x')} \tag{880}$$

D.2.3 Derivation 15 EE

Correct

$$\log(\phi_2) = C(\phi_2) \tag{881}$$

$$C(\phi_2) + \log(\phi_2) = 2C(\phi_2) \tag{882}$$

$$\frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) = \frac{d}{d\phi_2}2C(\phi_2) \tag{883}$$

$$\frac{d}{d\phi_2}C(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}C(\phi_2) \tag{884}$$

$$\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}\log(\phi_2) \tag{885}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = 4(\frac{d}{d\phi_2}\log(\phi_2))^2 \tag{886}$$

FLAN-T5

7.2 12.6 2.0 12.5
0.96 0.88 1.00 0.89

$$\log(\phi_2) = C(\phi_2) \tag{887}$$

$$\frac{d}{d\phi_2}\log(\phi_2) = \frac{d}{d\phi_2}C(\phi_2) \tag{888}$$

$$\frac{d}{d\phi_2}\log(\phi_2) + \frac{d}{d\phi_2}C(\phi_2) = 2\frac{d}{d\phi_2}C(\phi_2) \tag{889}$$

$$\frac{d}{d\phi_2}C(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}C(\phi_2) \tag{890}$$

$$\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}\log(\phi_2) \tag{891}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = 4(\frac{d}{d\phi_2}\log(\phi_2))^2 \tag{892}$$

GPT-4

-1.5 7.3 6.8 5.1
0.94 0.85 0.85 0.86

$$\log(\phi_2) = C(\phi_2) \tag{893}$$

$$\frac{d}{d\phi_2}\log(\phi_2) = \frac{d}{d\phi_2}C(\phi_2) \tag{894}$$

$$\frac{1}{\phi_2} = \frac{d}{d\phi_2}C(\phi_2) \tag{895}$$

$$\frac{d}{d\phi_2}C(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}C(\phi_2) \tag{896}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = 4(\frac{d}{d\phi_2}\log(\phi_2))^2 \tag{897}$$

ChatGPT

-11.5 -21.2 9.6 -21.5
0.82 0.61 0.86 0.68

$$\log(\phi_2) = C(\phi_2) \tag{898}$$

$$\frac{d}{d\phi_2}\log(\phi_2) = \frac{1}{\phi_2} \tag{899}$$

$$\frac{d}{d\phi_2}C(\phi_2) + \frac{1}{\phi_2} = 2\frac{d}{d\phi_2}C(\phi_2) \tag{900}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = (\frac{1}{\phi_2} + \frac{1}{\phi_2})^2 \tag{901}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = (\frac{2}{\phi_2})^2 \tag{902}$$

$$(\frac{d}{d\phi_2}\log(\phi_2) + \frac{1}{\phi_2})^2 = 4(\frac{d}{d\phi_2}\log(\phi_2))^2 \tag{903}$$

D.2.4 Derivation 16 EE

Correct

$$\hat{X}^t = \mathrm{v_t}\left(t, \hat{X}\right) \tag{904}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{905}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{906}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{907}$$

$$\hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{908}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{909}$$

FLAN-T5

-0.3 -9.6 10.5 -7.4
0.95 0.84 1.02 0.84

$$\hat{X}^t = \mathrm{v_t}\left(t, \hat{X}\right) \tag{910}$$

$$\hat{X}^t = \left(\mathrm{v_t}\left(t, \hat{X}\right)\right)^t \tag{911}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \frac{\partial}{\partial t}\left(\mathrm{v_t}\left(t, \hat{X}\right)\right)^t \tag{912}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \frac{\partial}{\partial t}\left(\mathrm{v_t}\left(t, \hat{X}\right)\right)^t \tag{913}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{914}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{915}$$

GPT-4

-8.3 -22.7 8.4 -18.7
0.90 0.76 0.98 0.76

$$\hat{X}^t = \mathrm{v_t}\left(t, \hat{X}\right) \tag{916}$$

$$\hat{X}^t \log\left(\hat{X}\right) = \log\left(\hat{X}\right) \mathrm{v_t}\left(t, \hat{X}\right) \tag{917}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \log\left(\hat{X}\right) \mathrm{v_t}\left(t, \hat{X}\right) \tag{918}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{919}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{920}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{921}$$

ChatGPT

14.2 32.7 17.5 24.6
0.80 0.63 0.74 0.68

$$\hat{X}^t = \mathrm{v_t}\left(t, \hat{X}\right) \tag{922}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{923}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) \tag{924}$$

$$\hat{X} + \hat{X}^t \log\left(\hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{925}$$

D.2.5 Derivation 17 EE

Correct

$$\frac{1}{A_x} = y(A_x) \tag{926}$$

$$\int \frac{1}{A_x} dA_x = \int y(A_x) dA_x \tag{927}$$

$$\varepsilon_0 + \log(A_x) = \int y(A_x) dA_x \tag{928}$$

$$\varepsilon_0 + \log(A_x) = \int \frac{1}{A_x} dA_x \tag{929}$$

$$\varepsilon_0 + \log(A_x) - \frac{x}{A_x} = \int \frac{1}{A_x} dA_x - \frac{x}{A_x} \tag{930}$$

$$\frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\int \frac{1}{A_x} dA_x - \frac{x}{A_x}) \tag{931}$$

FLAN-T5

6.7 7.3 4.9 7.3

0.96 0.94 0.98 0.94

$$\frac{1}{A_x} = y(A_x) \tag{932}$$

$$\int \frac{1}{A_x} dA_x = \int y(A_x) dA_x \tag{933}$$

$$\int \frac{1}{A_x} dA_x - \frac{x}{A_x} = \int y(A_x) dA_x \tag{934}$$

$$\varepsilon_0 + \log(A_x) = \int y(A_x) dA_x \tag{935}$$

$$\varepsilon_0 + \log(A_x) - \frac{x}{A_x} = \int \frac{1}{A_x} dA_x - \frac{x}{A_x} \tag{936}$$

$$\frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\int \frac{1}{A_x} dA_x - \frac{x}{A_x}) \tag{937}$$

GPT-4

-6.1 -23.4 -26.2 -17.6

0.88 0.75 0.91 0.75

$$\frac{1}{A_x} = y(A_x) \tag{938}$$

$$\int y(A_x) dA_x = \int \frac{1}{A_x} dA_x \tag{939}$$

$$\int \frac{1}{A_x} dA_x = \log(A_x) + C \tag{940}$$

$$\varepsilon_0 + \log(A_x) = \int y(A_x) dA_x \tag{941}$$

$$\varepsilon_0 + \log(A_x) = \log(A_x) + C \tag{942}$$

$$\frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\int \frac{1}{A_x} dA_x - \frac{x}{A_x}) \tag{943}$$

ChatGPT

0.3 0.1 -2.3 -0.4

0.92 0.89 0.94 0.90

$$\frac{1}{A_x} = y(A_x) \tag{944}$$

$$\log(A_x) = \int y(A_x) dA_x + \varepsilon_0 \tag{945}$$

$$\varepsilon_0 + \log(A_x) = \int y(A_x) dA_x \tag{946}$$

$$\frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\int \frac{1}{A_x} dA_x - \frac{x}{A_x}) \tag{947}$$

D.2.6 Derivation 18 EE

Correct

$$\sin(\lambda) = u(\lambda) \tag{948}$$

$$\int \sin(\lambda) d\lambda = \int u(\lambda) d\lambda \tag{949}$$

$$n - \cos(\lambda) = \int u(\lambda) d\lambda \tag{950}$$

$$n - \cos(\lambda) = \int \sin(\lambda) d\lambda \tag{951}$$

$$-\frac{n - \cos(\lambda)}{\cos(\lambda)} = -\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} \tag{952}$$

FLAN-T5

0.8 3.3 -2.2 3.4
0.88 0.87 0.88 0.88

$$\sin(\lambda) = u(\lambda) \tag{953}$$

$$\int \sin(\lambda) d\lambda = \int u(\lambda) d\lambda \tag{954}$$

$$\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = \frac{\int u(\lambda) d\lambda}{\cos(\lambda)} \tag{955}$$

$$n - \cos(\lambda) = \int u(\lambda) d\lambda \tag{956}$$

$$n - \cos(\lambda) = \int \sin(\lambda) d\lambda \tag{957}$$

$$-n - \cos(\lambda) = -\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} \tag{958}$$

GPT-4

2.7 2.1 7.0 1.9
0.91 0.94 0.87 0.95

$$\sin(\lambda) = u(\lambda) \tag{959}$$

$$n - \cos(\lambda) = \int u(\lambda) d\lambda \tag{960}$$

$$n - \cos(\lambda) = \int \sin(\lambda) d\lambda \tag{961}$$

$$-\frac{n - \cos(\lambda)}{\cos(\lambda)} = -\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} \tag{962}$$

ChatGPT

-5.7 -10.7 -4.5 -11.1
0.84 0.69 0.76 0.72

$$\sin(\lambda) = u(\lambda) \tag{963}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{964}$$

$$n - \cos(\lambda) = \int u(\lambda) d\lambda - \cos(\lambda) \tag{965}$$

$$-\frac{n - \cos(\lambda)}{\cos(\lambda)} = -\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} \tag{966}$$

D.2.7 Derivation 19 EE

Correct

$$\sin(\phi_1) = J(\phi_1) \tag{967}$$

$$\frac{d}{d\phi_1}\sin(\phi_1) = \frac{d}{d\phi_1}J(\phi_1) \tag{968}$$

$$\sin(\phi_1)\frac{d}{d\phi_1}\sin(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{969}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{970}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}\sin(\phi_1) \tag{971}$$

$$J(\phi_1)\cos(\phi_1) = J(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{972}$$

FLAN-T5

2.5 5.1 -3.0 4.2

0.93 0.93 0.94 0.93

$$\sin(\phi_1) = J(\phi_1) \tag{973}$$

$$\frac{d}{d\phi_1}\sin(\phi_1) = \frac{d}{d\phi_1}J(\phi_1) \tag{974}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{975}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{976}$$

$$J(\phi_1)\cos(\phi_1) = J(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{977}$$

GPT-4

-10.9 -31.1 -24.0 -22.4

0.79 0.67 1.22 0.67

$$\sin(\phi_1) = J(\phi_1) \tag{978}$$

$$\sin(\phi_1)\cos(\phi_1) = J(\phi_1)\cos(\phi_1) \tag{979}$$

$$\sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) = J(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{980}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{981}$$

$$J(\phi_1)\cos(\phi_1) = J(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{982}$$

ChatGPT

5.9 13.2 22.9 9.5

0.89 0.82 0.98 0.83

$$\sin(\phi_1) = J(\phi_1) \tag{983}$$

$$\cos(\phi_1) = \frac{d}{d\phi_1}\sin(\phi_1) \tag{984}$$

$$\sin(\phi_1)\cos(\phi_1) = \sin(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{985}$$

$$J(\phi_1) = \sin(\phi_1) \tag{986}$$

$$J(\phi_1)\cos(\phi_1) = J(\phi_1)\frac{d}{d\phi_1}J(\phi_1) \tag{987}$$

D.2.8 Derivation 20 EE

Correct

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{988}$$

$$\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{989}$$

$$\sin\left(\mathbf{J}_f\right)\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\sin\left(\mathbf{J}_f\right) \tag{990}$$

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{991}$$

$$\sin\left(\mathbf{J}_f\right)\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\sin\left(\mathbf{J}_f\right) \tag{992}$$

FLAN-T5

-4.8 -11.0 -19.8 -6.7
0.72 0.55 0.93 0.55

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{993}$$

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right)\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) \tag{994}$$

$$\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{995}$$

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right)\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right)\mathrm{f}'\left(\mathbf{J}_f\right) \tag{996}$$

$$\sin\left(\mathbf{J}_f\right)\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\sin\left(\mathbf{J}_f\right) \tag{997}$$

GPT-4

13.5 22.4 21.9 16.1
0.85 0.75 0.88 0.77

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{998}$$

$$\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{999}$$

$$\sin\left(\mathbf{J}_f\right)\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\sin\left(\mathbf{J}_f\right) \tag{1000}$$

ChatGPT

-15.2 -22.3 -20.3 -16.1
0.89 0.78 1.03 0.78

$$\frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{1001}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f}\sin\left(\mathbf{J}_f\right) \tag{1002}$$

$$\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right) \tag{1003}$$

$$\sin\left(\mathbf{J}_f\right)\cos\left(\mathbf{J}_f\right) = \mathrm{f}'\left(\mathbf{J}_f\right)\sin\left(\mathbf{J}_f\right) \tag{1004}$$

D.2.9 Derivation 21 EE

Correct

$$e^{\Psi_\lambda} = \mathbb{I}(\Psi_\lambda) \tag{1005}$$

$$\int e^{\Psi_\lambda} d\Psi_\lambda = \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1006}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1007}$$

$$\Psi_\lambda + \chi + e^{\Psi_\lambda} = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1008}$$

$$\Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1009}$$

FLAN-T5

1.9 2.7 1.1 2.7

0.98 0.97 1.09 0.97

$$\mathbb{I}(\Psi_\lambda) = \mathbb{I}(\Psi_\lambda) \tag{1010}$$

$$\int e^{\Psi_\lambda} d\Psi_\lambda = \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1011}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1012}$$

$$\Psi_\lambda + \chi + e^{\Psi_\lambda} = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1013}$$

$$\Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1014}$$

GPT-4

-10.4 -11.2 -8.8 -10.7

0.88 0.88 1.02 0.88

$$e^{\Psi_\lambda} = \mathbb{I}(\Psi_\lambda) \tag{1015}$$

$$\Psi_\lambda + \chi + e^{\Psi_\lambda} = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1016}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1017}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1018}$$

$$\Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1019}$$

ChatGPT

7.2 -5.9 -0.2 -6.1

0.85 0.79 0.81 0.80

$$e^{\Psi_\lambda} = \mathbb{I}(\Psi_\lambda) \tag{1020}$$

$$\frac{d}{d\Psi_\lambda} e^{\Psi_\lambda} = \frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) \tag{1021}$$

$$e^{\Psi_\lambda} = \mathbb{I}(\Psi_\lambda) \tag{1022}$$

$$\Psi_\lambda + \chi + e^{\Psi_\lambda} = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1023}$$

$$\frac{d}{d\Psi_\lambda} e^{\Psi_\lambda} = e^{\Psi_\lambda} = \mathbb{I}(\Psi_\lambda) \tag{1024}$$

$$\Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) = \Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda \tag{1025}$$

D.2.10 Derivation 22 EE

Correct

$$\log\left(\mathbf{s}\right) = \mathrm{y}'\left(\mathbf{s}\right) \tag{1026}$$

$$\int \log\left(\mathbf{s}\right) d\mathbf{s} = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1027}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1028}$$

$$\mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = a(\mathbf{s}) \tag{1029}$$

$$-\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) = a(\mathbf{s}) \tag{1030}$$

FLAN-T5

3.3 1.9 1.3 1.9
0.98 0.96 0.78 0.96

$$\log\left(\mathbf{s}\right) = \mathrm{y}'\left(\mathbf{s}\right) \tag{1031}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) = \mathbf{s}\,\mathrm{y}'\left(\mathbf{s}\right) \tag{1032}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1033}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1034}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1035}$$

$$\mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = a(\mathbf{s}) \tag{1036}$$

$$-\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) = a(\mathbf{s}) \tag{1037}$$

GPT-4

12.4 21.3 22.6 19.5
0.87 0.79 0.76 0.80

$$\log\left(\mathbf{s}\right) = \mathrm{y}'\left(\mathbf{s}\right) \tag{1038}$$

$$\mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = a(\mathbf{s}) \tag{1039}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1040}$$

$$-\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) = a(\mathbf{s}) \tag{1041}$$

ChatGPT

0.2 2.1 -4.8 2.1
0.92 0.93 0.73 0.93

$$\log\left(\mathbf{s}\right) = \mathrm{y}'\left(\mathbf{s}\right) \tag{1042}$$

$$\frac{d}{d\mathbf{s}}\log\left(\mathbf{s}\right) = \frac{d}{d\mathbf{s}}\mathrm{y}'\left(\mathbf{s}\right) \tag{1043}$$

$$\frac{1}{\mathbf{s}} = \mathrm{y}''\left(\mathbf{s}\right) \tag{1044}$$

$$\int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = \log\left(\mathbf{s}\right) + \omega \tag{1045}$$

$$\mathrm{y}'\left(\mathbf{s}\right) - \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} = a(\mathbf{s}) \tag{1046}$$

$$\mathbf{s}\log\left(\mathbf{s}\right) - \mathbf{s} + \omega = \int \mathrm{y}'\left(\mathbf{s}\right) d\mathbf{s} \tag{1047}$$

$$-\mathbf{s}\log\left(\mathbf{s}\right) + \mathbf{s} - \omega + \mathrm{y}'\left(\mathbf{s}\right) = a(\mathbf{s}) \tag{1048}$$

D.2.11 Derivation 23 EE

Correct

$$\sin\left(F_{N}\right) = \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) \tag{1049}$$

$$\int \sin\left(F_{N}\right) dF_{N} = \int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N} \tag{1050}$$

$$(\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1051}$$

$$(\int \sin\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1052}$$

$$(Q - \cos\left(F_{N}\right))^{2} = \mathbf{v}(F_{N}) \tag{1053}$$

$$(\int \sin\left(F_{N}\right) dF_{N})^{2} = (\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} \tag{1054}$$

$$(Q - \cos\left(F_{N}\right))^{2} = (\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} \tag{1055}$$

$$(Q - \cos\left(F_{N}\right))^{2} = (\int \sin\left(F_{N}\right) dF_{N})^{2} \tag{1056}$$

FLAN-T5

-8.6 -22.6 8.6 -18.0
0.82 0.75 0.95 0.75

$$\sin\left(F_{N}\right) = \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) \tag{1057}$$

$$\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N} = \int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N} \tag{1058}$$

$$(\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} = (\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} \tag{1059}$$

$$(\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1060}$$

$$(\int \sin\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1061}$$

$$(Q - \cos\left(F_{N}\right))^{2} = \mathbf{v}(F_{N}) \tag{1062}$$

$$(Q - \cos\left(F_{N}\right))^{2} = (\int \sin\left(F_{N}\right) dF_{N})^{2} \tag{1063}$$

GPT-4

7.2 12.8 12.1 8.9
0.87 0.82 0.85 0.83

$$\sin\left(F_{N}\right) = \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) \tag{1064}$$

$$(\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1065}$$

$$(\int \sin\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1066}$$

$$(Q - \cos\left(F_{N}\right))^{2} = \mathbf{v}(F_{N}) \tag{1067}$$

$$(Q - \cos\left(F_{N}\right))^{2} = (\int \sin\left(F_{N}\right) dF_{N})^{2} \tag{1068}$$

ChatGPT

-14.5 -26.0 -0.1 -21.7
0.69 0.70 0.79 0.70

$$\sin\left(F_{N}\right) = \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) \tag{1069}$$

$$\cos\left(F_{N}\right) = \sqrt{1 - \sin^{2}\left(F_{N}\right)} \tag{1070}$$

$$Q = \cos\left(F_{N}\right) + \int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N} \tag{1071}$$

$$(Q - \cos\left(F_{N}\right))^{2} = (\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} \tag{1072}$$

$$(\int \mathrm{A}_{\mathrm{z}}\left(F_{N}\right) dF_{N})^{2} = \mathbf{v}(F_{N}) \tag{1073}$$

$$(Q - \cos\left(F_{N}\right))^{2} = \mathbf{v}(F_{N}) \tag{1074}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \mathbf{v}(F_N) \tag{1075}$$

$$(Q - \cos\left(F_N\right))^2 = (\int \sin\left(F_N\right) dF_N)^2 \tag{1076}$$

D.2.12 Derivation 24 EE

Correct

$$\sin\left(\varepsilon_0\right) = \mathrm{f}'\left(\varepsilon_0\right) \tag{1077}$$

$$-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right) = 0 \tag{1078}$$

$$\frac{d}{d\varepsilon_0}(-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right)) = \frac{d}{d\varepsilon_0} 0 \tag{1079}$$

$$\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = 0 \tag{1080}$$

$$\int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 = \int 0 d\varepsilon_0 \tag{1081}$$

FLAN-T5

12.2 19.5 21.6 19.4
0.93 0.84 0.87 0.84

$$\sin\left(\varepsilon_0\right) = \mathrm{f}'\left(\varepsilon_0\right) \tag{1082}$$

$$\frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1083}$$

$$\cos\left(\varepsilon_0\right) + \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) + \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1084}$$

$$\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = 0 \tag{1085}$$

$$\int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 = \int 0 d\varepsilon_0 \tag{1086}$$

GPT-4

-6.4 -16.5 -11.1 -14.2
0.86 0.78 1.09 0.78

$$\sin\left(\varepsilon_0\right) = \mathrm{f}'\left(\varepsilon_0\right) \tag{1087}$$

$$\frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1088}$$

$$\cos\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1089}$$

$$\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = 0 \tag{1090}$$

$$\int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 = \int 0 d\varepsilon_0 \tag{1091}$$

ChatGPT

2.4 4.0 13.0 3.8
0.92 0.92 1.00 0.92

$$\sin\left(\varepsilon_0\right) = \mathrm{f}'\left(\varepsilon_0\right) \tag{1092}$$

$$\cos\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1093}$$

$$\cos\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1094}$$

$$\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = 0 \tag{1095}$$

$$\int (\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right)) d\varepsilon_0 = \int 0 d\varepsilon_0 \tag{1096}$$

D.2.13 Derivation 25 EE

Correct

$$W + \frac{q}{B} = y(W, q, B) \tag{1097}$$

$$W - y(W, q, B) + \frac{q}{B} = 0 \tag{1098}$$

$$\frac{\partial}{\partial q}(W - y(W, q, B) + \frac{q}{B}) = \frac{d}{dq}0 \tag{1099}$$

$$-\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} = 0 \tag{1100}$$

$$-\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} = 0 \tag{1101}$$

FLAN-T5

6.8 7.4 9.4 7.2
0.85 0.91 1.07 0.91

$$y(W, q, B) = W + \frac{q}{B} \tag{1102}$$

$$y(W, q, B) - \frac{1}{B} = 0 \tag{1103}$$

$$\frac{\partial}{\partial q}(y(W, q, B) - \frac{1}{B}) = \frac{d}{dq}0 \tag{1104}$$

$$-\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} = 0 \tag{1105}$$

$$-\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} = 0 \tag{1106}$$

GPT-4

9.0 12.2 5.1 12.1
1.10 0.93 1.17 0.93

$$W + \frac{q}{B} = y(W, q, B) \tag{1107}$$

$$y(W, q, B) = W + \frac{q}{B} \tag{1108}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{\partial}{\partial q}(W + \frac{q}{B}) \tag{1109}$$

$$-\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{1110}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{1111}$$

ChatGPT

2.3 11.0 11.0 8.8
0.96 0.79 0.97 0.81

$$W + \frac{q}{B} = y(W, q, B) \tag{1112}$$

$$\frac{\partial}{\partial q}(W + \frac{q}{B}) = \frac{1}{B} \tag{1113}$$

$$-\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} = 0 \tag{1114}$$

$$-\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} = 0 \tag{1115}$$

D.3 Alternative Goal (AG)

D.3.1 Derivation 13 AG

Correct

$$W(P_e) = \log(P_e) \tag{1116}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1117}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{1118}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{1119}$$

$$-P_e + \frac{d}{dP_e}\log(P_e) = -P_e + \frac{1}{P_e} \tag{1120}$$

FLAN-T5

6.0 8.4 -9.2 8.4

0.92 0.90 0.96 0.90

$$(P_e) = \log(P_e) \tag{1121}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1122}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{1123}$$

$$-P_e + \frac{d}{dP_e}W(P_e) = -P_e + \frac{1}{P_e} \tag{1124}$$

$$-P_e + \frac{d}{dP_e}\log(P_e) = -P_e + \frac{1}{P_e} \tag{1125}$$

GPT-4

0.4 0.0 -10.0 0.0

0.97 0.98 0.86 0.98

$$W(P_e) = \log(P_e) \tag{1126}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1127}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{1128}$$

$$-P_e + \frac{d}{dP_e}\log(P_e) = -P_e + \frac{1}{P_e} \tag{1129}$$

ChatGPT

-10.7 -21.1 -24.2 -19.3

0.86 0.80 0.99 0.80

$$W(P_e) = \log(P_e) \tag{1130}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1131}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{1132}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{1133}$$

$$-P_e + \frac{d}{dP_e}\log(P_e) = -P_e + \frac{1}{P_e} \tag{1134}$$

D.3.2 Derivation 14 AG

Correct

$$\phi(x') = \int \log(x')dx' \tag{1135}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{1136}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1137}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1138}$$

$$\mathrm{t}_1(x', n_2) = \frac{d}{dx'}\phi(x') \tag{1139}$$

$$\frac{\mathrm{t}_1(x', n_2)}{n_2 + x'\log(x') - x'} = \frac{\frac{d}{dx'}\phi(x')}{n_2 + x'\log(x') - x'} \tag{1140}$$

FLAN-T5

4.4 2.5 -5.6 3.3
0.86 0.74 0.98 0.74

$$\phi(x') = \int \log(x')dx' \tag{1141}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{1142}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1143}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1144}$$

$$\frac{\mathrm{t}_1(x', n_2)}{n_2 + x'\log(x') - x'} = \frac{\frac{d}{dx'}\int \log(x')dx'}{n_2 + x'\log(x') - x'} \tag{1145}$$

$$\frac{\mathrm{t}_1(x', n_2)}{n_2 + x'\log(x') - x'} = \frac{\frac{d}{dx'}\phi(x')}{n_2 + x'\log(x') - x'} \tag{1146}$$

GPT-4

-5.9 -12.8 0.0 -12.9
0.93 0.82 0.96 0.84

$$\phi(x') = \int \log(x')dx' \tag{1147}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{1148}$$

$$\frac{d}{dx'}\phi(x') = \log(x') \tag{1149}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1150}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1151}$$

$$\frac{\mathrm{t}_1(x', n_2)}{n_2 + x'\log(x') - x'} = \frac{\frac{d}{dx'}\phi(x')}{n_2 + x'\log(x') - x'} \tag{1152}$$

ChatGPT

-14.4 -27.9 9.8 -28.0
0.74 0.55 0.74 0.61

$$\phi(x') = \int \log(x')dx' \tag{1153}$$

$$\frac{d}{dx'}\phi(x') = \frac{1}{x'} \tag{1154}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1155}$$

$$\frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') = \log(x') \tag{1156}$$

$$\frac{\frac{d}{dx'}\phi(x')}{n_2 + x'\log(x') - x'} = \frac{1}{x'} \tag{1157}$$

$$\frac{\mathrm{t}_1(x', n_2)}{n_2 + x'\log(x') - x'} = \frac{\frac{d}{dx'}\phi(x')}{n_2 + x'\log(x') - x'} \tag{1158}$$

D.3.3 Derivation 15 AG

Correct

$$C(\phi_2) = \log(\phi_2) \tag{1159}$$

$$2C(\phi_2) = C(\phi_2) + \log(\phi_2) \tag{1160}$$

$$\frac{d}{d\phi_2} 2C(\phi_2) = \frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) \tag{1161}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{1162}$$

$$2\frac{d}{d\phi_2} \log(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1163}$$

$$\phi_2 + 2\frac{d}{d\phi_2} \log(\phi_2) = \phi_2 + \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1164}$$

FLAN-T5

-0.0 0.0 0.1 0.0

1.00 1.00 0.98 1.00

$$(\phi_2) = \log(\phi_2) \tag{1165}$$

$$2C(\phi_2) = C(\phi_2) + \log(\phi_2) \tag{1166}$$

$$\frac{d}{d\phi_2} 2C(\phi_2) = \frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) \tag{1167}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{1168}$$

$$2\frac{d}{d\phi_2} \log(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1169}$$

$$\phi_2 + 2\frac{d}{d\phi_2} \log(\phi_2) = \phi_2 + \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1170}$$

GPT-4

1.5 12.9 -1.0 9.9

0.92 0.80 0.80 0.82

$$C(\phi_2) = \log(\phi_2) \tag{1171}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{1}{\phi_2} \tag{1172}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = 2\frac{1}{\phi_2} \tag{1173}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{1174}$$

$$\phi_2 + 2\frac{d}{d\phi_2} \log(\phi_2) = \phi_2 + \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1175}$$

ChatGPT

-18.1 -29.4 -10.8 -29.5

0.82 0.58 0.96 0.65

$$C(\phi_2) = \log(\phi_2) \tag{1176}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) \tag{1177}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = 2\frac{d}{d\phi_2} \log(\phi_2) \tag{1178}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{d}{d\phi_2} \log(\phi_2) \tag{1179}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{1180}$$

$$\phi_2 + 2\frac{d}{d\phi_2} \log(\phi_2) = \phi_2 + \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1181}$$

D.3.4 Derivation 16 AG

Correct

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1182}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1183}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{1184}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1185}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right) \tag{1186}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right))^t \tag{1187}$$

FLAN-T5

-7.5 -20.3 -0.3 -17.8
0.90 0.78 0.98 0.78

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1188}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1189}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{1190}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1191}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X} + \hat{X}^t \log\left(\hat{X}\right))^t \tag{1192}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right))^t \tag{1193}$$

GPT-4

-4.2 -8.0 -0.5 -6.5
0.96 0.90 0.97 0.90

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1194}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1195}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \log\left(\hat{X}\right) \tag{1196}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1197}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right))^t \tag{1198}$$

ChatGPT

0.2 -4.8 1.0 -4.2
0.94 0.87 0.87 0.87

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1199}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1200}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \hat{X}^t \log\left(\hat{X}\right) \tag{1201}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1202}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X}^t + \hat{X}^t \log\left(\hat{X}\right))^t \tag{1203}$$

$$(\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right))^t = (\hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right))^t \tag{1204}$$

D.3.5 Derivation 17 AG

Correct

$$y(A_x) = \frac{1}{A_x} \tag{1205}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1206}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log(A_x) \tag{1207}$$

$$\int \frac{1}{A_x}dA_x = \varepsilon_0 + \log(A_x) \tag{1208}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \varepsilon_0 + \log(A_x) - \frac{x}{A_x} \tag{1209}$$

$$\frac{\partial}{\partial \varepsilon_0}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial \varepsilon_0}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1210}$$

FLAN-T5

0.0 0.0 4.9 0.0
1.00 1.00 1.00 1.00

$$y(A_x) = \frac{1}{A_x} \tag{1211}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1212}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log(A_x) \tag{1213}$$

$$\int \frac{1}{A_x}dA_x = \varepsilon_0 + \log(A_x) \tag{1214}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \varepsilon_0 + \log(A_x) - \frac{x}{A_x} \tag{1215}$$

$$\frac{\partial}{\partial \varepsilon_0}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial \varepsilon_0}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1216}$$

GPT-4

-0.5 -1.0 -12.6 -0.7
0.97 0.98 0.84 0.98

$$y(A_x) = \frac{1}{A_x} \tag{1217}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1218}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log(A_x) \tag{1219}$$

$$\frac{\partial}{\partial \varepsilon_0}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial \varepsilon_0}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1220}$$

ChatGPT

-8.0 -18.0 -26.1 -11.8
0.64 0.48 0.92 0.48

$$y(A_x) = \frac{1}{A_x} \tag{1221}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1222}$$

$$\int y(A_x)dA_x = \log(A_x) + q \tag{1223}$$

$$\int \frac{1}{A_x}dA_x = \log(A_x) + q \tag{1224}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \log(A_x) - \frac{x}{A_x} + q \tag{1225}$$

$$\frac{\partial}{\partial \varepsilon_0}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial \varepsilon_0}(\log(A_x) - \frac{x}{A_x} + \varepsilon_0) \tag{1226}$$

$$\frac{\partial}{\partial \varepsilon_0}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial \varepsilon_0}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1227}$$

D.3.6 Derivation 18 AG

Correct

$$u(\lambda) = \sin(\lambda) \tag{1228}$$

$$\int u(\lambda)d\lambda = \int \sin(\lambda)d\lambda \tag{1229}$$

$$\int u(\lambda)d\lambda = n - \cos(\lambda) \tag{1230}$$

$$\int \sin(\lambda)d\lambda = n - \cos(\lambda) \tag{1231}$$

$$\iint \sin(\lambda)d\lambda dn = \int (n - \cos(\lambda))dn \tag{1232}$$

FLAN-T5

-14.0 -20.8 -4.8 -20.8
0.86 0.75 0.45 0.79

$$u(\lambda) = \sin(\lambda) \tag{1233}$$

$$\int u(\lambda)d\lambda = \int \sin(\lambda)d\lambda \tag{1234}$$

$$\int u(\lambda)d\lambda = n - \cos(\lambda) \tag{1235}$$

$$\int \sin(\lambda)d\lambda = n - \cos(\lambda) \tag{1236}$$

$$\iint \sin(\lambda)d\lambda dn = \int (n - \cos(\lambda))dn \tag{1237}$$

GPT-4

0.8 -12.3 7.5 -11.0
0.91 1.07 1.02 1.06

$$u(\lambda) = \sin(\lambda) \tag{1238}$$

$$\int u(\lambda)d\lambda = \int \sin(\lambda)d\lambda \tag{1239}$$

$$\int u(\lambda)d\lambda = -\cos(\lambda) + C \tag{1240}$$

$$\int u(\lambda)d\lambda = n - \cos(\lambda) \tag{1241}$$

$$\iint \sin(\lambda)d\lambda dn = \int (n - \cos(\lambda))dn \tag{1242}$$

ChatGPT

13.7 16.7 19.1 16.7
0.68 0.86 0.74 0.86

$$u(\lambda) = \sin(\lambda) \tag{1243}$$

$$\int u(\lambda)d\lambda = \int \sin(\lambda)d\lambda \tag{1244}$$

$$\int \sin(\lambda)d\lambda = -\cos(\lambda) + C \tag{1245}$$

$$\int u(\lambda)d\lambda = -\cos(\lambda) + C \tag{1246}$$

$$\int_0^n \int_0^\lambda \sin(\lambda)d\lambda dn = \int_0^n (-\cos(\lambda))dn + \int_0^n C dn \tag{1247}$$

$$\iint \sin(\lambda)d\lambda dn = \int (n - \cos(\lambda))dn \tag{1248}$$

D.3.7 Derivation 19 AG

Correct

$$J(\phi_1) = \sin(\phi_1) \tag{1249}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{1250}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{1251}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1252}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1253}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) \tag{1254}$$

FLAN-T5

-10.2 -20.7 2.7 -18.9

0.89 0.81 0.97 0.81

$$(\phi_1) = \sin(\phi_1) \tag{1255}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{1256}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{1257}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1258}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1259}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) \tag{1260}$$

GPT-4

-5.3 -9.2 -22.0 -6.5

0.92 0.88 0.98 0.88

$$J(\phi_1) = \sin(\phi_1) \tag{1261}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{1262}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1263}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) \tag{1264}$$

ChatGPT

-9.0 -17.4 -10.1 -14.4

0.87 0.82 0.96 0.82

$$J(\phi_1) = \sin(\phi_1) \tag{1265}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{1266}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1267}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1268}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{1269}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) - \frac{d}{d\phi_1} J(\phi_1) \tag{1270}$$

D.3.8 Derivation 20 AG

Correct

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1271}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1272}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1273}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1274}$$

$$\frac{d}{d\mathbf{J}_f} \cos\left(\mathbf{J}_f\right) = \frac{d^2}{d\mathbf{J}_f^2} \sin\left(\mathbf{J}_f\right) \tag{1275}$$

FLAN-T5

-3.7 -1.1 -5.3 -1.0
0.94 0.96 0.90 0.96

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1276}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1277}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1278}$$

$$\frac{d}{d\mathbf{J}_f} \cos\left(\mathbf{J}_f\right) = \frac{d^2}{d\mathbf{J}_f^2} \sin\left(\mathbf{J}_f\right) \tag{1279}$$

GPT-4

11.7 21.4 19.5 15.5
0.86 0.77 0.84 0.79

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1280}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1281}$$

$$\frac{d}{d\mathbf{J}_f} \cos\left(\mathbf{J}_f\right) = \frac{d^2}{d\mathbf{J}_f^2} \sin\left(\mathbf{J}_f\right) \tag{1282}$$

ChatGPT

-17.5 -23.4 -18.7 -17.1
0.82 0.77 0.82 0.77

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1283}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1284}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1285}$$

$$\frac{d}{d\mathbf{J}_f} \cos\left(\mathbf{J}_f\right) = \frac{d^2}{d\mathbf{J}_f^2} \sin\left(\mathbf{J}_f\right) \tag{1286}$$

D.3.9 Derivation 21 AG

Correct

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1287}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1288}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1289}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1290}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1291}$$

FLAN-T5

0.0 0.0 -0.9 0.0
1.00 1.00 1.05 1.00

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1292}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1293}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1294}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1295}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1296}$$

GPT-4

-3.8 -0.9 -2.1 -0.9
0.93 0.95 0.89 0.95

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1297}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1298}$$

$$\int e^{\Psi_\lambda} d\Psi_\lambda = \chi + e^{\Psi_\lambda} \tag{1299}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1300}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1301}$$

ChatGPT

-3.5 -8.2 0.3 -8.2
0.95 0.89 1.03 0.90

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1302}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = \frac{d}{d\Psi_\lambda} e^{\Psi_\lambda} \tag{1303}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1304}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1305}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1306}$$

$$\Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1307}$$

D.3.10 Derivation 22 AG

Correct

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1308}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \int \log(\mathbf{s})d\mathbf{s} \tag{1309}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1310}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s})d\mathbf{s} \tag{1311}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \log(\mathbf{s})d\mathbf{s} \tag{1312}$$

FLAN-T5

-7.1 -20.0 -6.8 -20.0
0.92 0.74 1.00 0.77

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1313}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \int \log(\mathbf{s})d\mathbf{s} \tag{1314}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1315}$$

$$\mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1316}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \log(\mathbf{s})d\mathbf{s} \tag{1317}$$

GPT-4

0.0 0.2 3.2 0.2
1.00 1.00 1.00 1.00

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1318}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s})d\mathbf{s} \tag{1319}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \int \log(\mathbf{s})d\mathbf{s} \tag{1320}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1321}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \log(\mathbf{s})d\mathbf{s} \tag{1322}$$

ChatGPT

5.4 9.2 -5.6 9.2
0.99 0.93 0.98 0.93

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1323}$$

$$\frac{d}{d\mathbf{s}}\mathrm{y}'(\mathbf{s}) = \frac{d}{d\mathbf{s}}\log(\mathbf{s}) \tag{1324}$$

$$\int \frac{d}{d\mathbf{s}}\mathrm{y}'(\mathbf{s})d\mathbf{s} = \int \frac{d}{d\mathbf{s}}\log(\mathbf{s})d\mathbf{s} \tag{1325}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s})d\mathbf{s} \tag{1326}$$

$$\int \mathrm{y}'(\mathbf{s})d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1327}$$

$$\int \log(\mathbf{s})d\mathbf{s} = \log(\mathbf{s})\mathbf{s} - \mathbf{s} \tag{1328}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \log(\mathbf{s})d\mathbf{s} \tag{1329}$$

D.3.11 Derivation 23 AG

Correct

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1330}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{1331}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1332}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1333}$$

$$\mathbf{v}(F_N) = (Q - \cos\left(F_N\right))^2 \tag{1334}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = (\int \sin\left(F_N\right) dF_N)^2 \tag{1335}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = (Q - \cos\left(F_N\right))^2 \tag{1336}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = (Q - \cos\left(F_N\right))^2 \tag{1337}$$

FLAN-T5

0.0 0.0 0.0 0.0
1.00 1.00 0.98 1.00

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1338}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{1339}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1340}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1341}$$

$$\mathbf{v}(F_N) = (Q - \cos\left(F_N\right))^2 \tag{1342}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = (Q - \cos\left(F_N\right))^2 \tag{1343}$$

GPT-4

7.2 12.7 13.3 8.8
0.91 0.85 0.86 0.86

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1344}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1345}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1346}$$

$$\mathbf{v}(F_N) = (Q - \cos\left(F_N\right))^2 \tag{1347}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = (Q - \cos\left(F_N\right))^2 \tag{1348}$$

ChatGPT

0.0 0.0 0.0 0.0
1.00 1.00 1.03 1.00

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1349}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1350}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1351}$$

$$\mathbf{v}(F_N) = (\frac{-\cos\left(F_N\right) + Q}{\sqrt{2}})^2 \tag{1352}$$

$$\mathbf{v}(F_N) = (Q - \cos\left(F_N\right))^2 \tag{1353}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = (Q - \cos\left(F_N\right))^2 \tag{1354}$$

D.3.12 Derivation 24 AG

Correct

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{1355}$$

$$0 = -\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right) \tag{1356}$$

$$\frac{d}{d\varepsilon_0} 0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right)) \tag{1357}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1358}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1359}$$

FLAN-T5

0.0 0.1 -0.5 0.1
1.00 1.00 1.00 1.00

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{1360}$$

$$0 = -\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right) \tag{1361}$$

$$\frac{d}{d\varepsilon_0} 0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'\left(\varepsilon_0\right) + \sin\left(\varepsilon_0\right)) \tag{1362}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1363}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1364}$$

GPT-4

-5.0 -14.5 -15.6 -12.2
0.90 0.92 0.87 0.91

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{1365}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1366}$$

$$\frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) \tag{1367}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1368}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1369}$$

ChatGPT

9.2 16.6 3.2 16.6
0.95 0.96 0.91 0.95

$$\mathrm{f}'\left(\varepsilon_0\right) = \sin\left(\varepsilon_0\right) \tag{1370}$$

$$\frac{d}{d\varepsilon_0}\cos\left(\varepsilon_0\right) = -\sin\left(\varepsilon_0\right) \tag{1371}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) \tag{1372}$$

$$\cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) = \cos\left(\varepsilon_0\right) - \cos\left(\varepsilon_0\right) \tag{1373}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\mathrm{f}'\left(\varepsilon_0\right) \tag{1374}$$

$$0 = \cos\left(\varepsilon_0\right) - \frac{d}{d\varepsilon_0}\sin\left(\varepsilon_0\right) \tag{1375}$$

D.3.13 Derivation 25 AG

Correct

$$y(W, q, B) = W + \frac{q}{B} \tag{1376}$$

$$0 = W - y(W, q, B) + \frac{q}{B} \tag{1377}$$

$$\frac{d}{dq}0 = \frac{\partial}{\partial q}(W - y(W, q, B) + \frac{q}{B}) \tag{1378}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1379}$$

$$W + \frac{q}{B} = W - \frac{\partial}{\partial q}y(W, q, B) + \frac{q}{B} + \frac{1}{B} \tag{1380}$$

FLAN-T5

2.2 11.3 -6.4 11.4

0.89 0.82 0.99 0.82

$$y(W, q, B) = W + \frac{q}{B} \tag{1381}$$

$$0 = W - y(W, q, B) + \frac{q}{B} \tag{1382}$$

$$\frac{d}{dq}0 = \frac{\partial}{\partial q}(W - y(W, q, B) + \frac{q}{B}) \tag{1383}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1384}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1385}$$

$$W + \frac{q}{B} = W - \frac{\partial}{\partial q}y(W, q, B) + \frac{q}{B} + \frac{1}{B} \tag{1386}$$

GPT-4

-0.4 10.6 7.8 8.5

0.98 0.89 0.92 0.91

$$y(W, q, B) = W + \frac{q}{B} \tag{1387}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{1}{B} \tag{1388}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1389}$$

$$W + \frac{q}{B} = W - \frac{\partial}{\partial q}y(W, q, B) + \frac{q}{B} + \frac{1}{B} \tag{1390}$$

ChatGPT

-3.0 -3.4 -3.9 -3.3

0.91 0.92 0.90 0.93

$$y(W, q, B) = W + \frac{q}{B} \tag{1391}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{1}{B} \tag{1392}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1393}$$

$$-\frac{\partial}{\partial q}y(W, q, B) = -\frac{q}{B^2} \tag{1394}$$

$$W + \frac{q}{B} = W - \frac{\partial}{\partial q}y(W, q, B) + \frac{q}{B} + \frac{1}{B} \tag{1395}$$

D.4 Step Removal (SR)

D.4.1 Derivation 13 SR

Correct

$$W(P_e) = \log(P_e) \tag{1396}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1397}$$

$$\frac{d}{dP_e}W(P_e) = \frac{1}{P_e} \tag{1398}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{1399}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{1400}$$

FLAN-T5

14.6 12.2 3.4 12.0
0.85 0.88 0.86 0.88

$$(P_e) = \log(P_e) \tag{1401}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1402}$$

$$\int \frac{d}{dP_e}W(P_e)dP_e = \int \frac{d}{dP_e}\log(P_e)dP_e \tag{1403}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{1404}$$

GPT-4

0.0 -0.6 -1.0 -0.5
0.96 0.98 0.89 0.98

$$W(P_e) = \log(P_e) \tag{1405}$$

$$\frac{d}{dP_e}W(P_e) = \frac{d}{dP_e}\log(P_e) \tag{1406}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{1407}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{1408}$$

ChatGPT

18.1 32.5 13.3 24.4
0.81 0.65 0.65 0.70

$$W(P_e) = \log(P_e) \tag{1409}$$

$$\frac{d}{dP_e}\log(P_e) = \frac{1}{P_e} \tag{1410}$$

$$\int \frac{d}{dP_e}\log(P_e)dP_e = \int \frac{1}{P_e}dP_e \tag{1411}$$

D.4.2 Derivation 14 SR

Correct

$$\phi(x') = \int \log(x')dx' \tag{1412}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{1413}$$

$$\frac{d}{dx'}\phi(x') = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1414}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1415}$$

$$\mathrm{t}_1(x', n_2) = \frac{d}{dx'}\phi(x') \tag{1416}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{1417}$$

FLAN-T5

3.4 6.0 20.0 5.1
0.92 0.90 0.73 0.90

$$\phi(x') = \int \log(x')dx' \tag{1418}$$

$$\frac{d}{dx'}\phi(x') = \frac{d}{dx'}\int \log(x')dx' \tag{1419}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1420}$$

$$\mathrm{t}_1(x', n_2) = \frac{d}{dx'}\phi(x') \tag{1421}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{1422}$$

GPT-4

1.3 -2.1 16.5 -2.1
0.89 0.87 0.50 0.87

$$\phi(x') = \int \log(x')dx' \tag{1423}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1424}$$

$$\frac{d}{dx'}\phi(x') = \log(x') \tag{1425}$$

$$e^{-\frac{d}{dx'}\phi(x')} = e^{-\log(x')} \tag{1426}$$

$$e^{-\log(x')} = \frac{1}{x'} \tag{1427}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x')\frac{1}{x'} \tag{1428}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{1429}$$

ChatGPT

-14.1 -23.5 9.6 -23.5
0.79 0.65 0.72 0.69

$$\phi(x') = \int \log(x')dx' \tag{1430}$$

$$\frac{d}{dx'}\phi(x') = \frac{1}{x'} \tag{1431}$$

$$\mathrm{t}_1(x', n_2) = \frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1432}$$

$$\mathrm{t}_1(x', n_2) = \log(x') + 1 - \frac{n_2}{x'} \tag{1433}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{\partial}{\partial x'}(n_2 + x'\log(x') - x') \tag{1434}$$

$$\mathrm{t}_1(x', n_2)e^{-\frac{d}{dx'}\phi(x')} = e^{-\frac{d}{dx'}\phi(x')}\frac{d}{dx'}\phi(x') \tag{1435}$$

D.4.3 Derivation 15 SR

Correct

$$C(\phi_2) = \log(\phi_2) \tag{1436}$$

$$2C(\phi_2) = C(\phi_2) + \log(\phi_2) \tag{1437}$$

$$\frac{d}{d\phi_2} 2C(\phi_2) = \frac{d}{d\phi_2}(C(\phi_2) + \log(\phi_2)) \tag{1438}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{1}{\phi_2} \tag{1439}$$

$$2\frac{d}{d\phi_2} \log(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2} \tag{1440}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{1441}$$

FLAN-T5

11.3 14.5 19.2 14.2
0.89 0.85 0.73 0.86

$$(\phi_2) = \log(\phi_2) \tag{1442}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) \tag{1443}$$

$$2\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} C(\phi_2) + \frac{d}{d\phi_2} \log(\phi_2) \tag{1444}$$

$$4(\frac{d}{d\phi_2} C(\phi_2))^2 = (\frac{d}{d\phi_2} C(\phi_2) + \frac{d}{d\phi_2} \log(\phi_2))^2 \tag{1445}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{1446}$$

GPT-4

12.7 9.6 12.6 9.0
0.85 0.86 0.84 0.86

$$C(\phi_2) = \log(\phi_2) \tag{1447}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) \tag{1448}$$

$$\frac{d}{d\phi_2} \log(\phi_2) = \frac{1}{\phi_2} \tag{1449}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = 4(\frac{1}{\phi_2})^2 \tag{1450}$$

$$(\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 = (\frac{1}{\phi_2} + \frac{1}{\phi_2})^2 \tag{1451}$$

$$4(\frac{1}{\phi_2})^2 = (\frac{2}{\phi_2})^2 \tag{1452}$$

ChatGPT

-8.7 -16.6 6.5 -16.7
0.81 0.68 0.64 0.72

$$C(\phi_2) = \log(\phi_2) \tag{1453}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{d}{d\phi_2} \log(\phi_2) \tag{1454}$$

$$\frac{d}{d\phi_2} C(\phi_2) = \frac{1}{\phi_2} \tag{1455}$$

$$(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{1}{\phi_2})^2 \tag{1456}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = 4(\frac{1}{\phi_2})^2 \tag{1457}$$

$$(\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 = (\frac{1}{\phi_2} + \frac{d}{d\phi_2} \log(\phi_2))^2 \tag{1458}$$

$$4(\frac{d}{d\phi_2} \log(\phi_2))^2 = (\frac{d}{d\phi_2} \log(\phi_2) + \frac{1}{\phi_2})^2 \tag{1459}$$

D.4.4 Derivation 16 SR

Correct

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1460}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1461}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{1462}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1463}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \mathrm{v_t}\left(t, \hat{X}\right) \log\left(\hat{X}\right) \tag{1464}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1465}$$

FLAN-T5

0.0 0.0 0.0 0.0
1.00 1.00 0.99 1.00

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1466}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1467}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \frac{\partial}{\partial t}\hat{X}^t \tag{1468}$$

$$\hat{X} + \frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1469}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1470}$$

GPT-4

12.5 25.7 47.6 18.2
0.85 0.70 0.77 0.73

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1471}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1472}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \hat{X}^t \log\left(\hat{X}\right) \tag{1473}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1474}$$

ChatGPT

20.6 42.8 47.6 31.8
0.74 0.51 0.59 0.60

$$\mathrm{v_t}\left(t, \hat{X}\right) = \hat{X}^t \tag{1475}$$

$$\frac{\partial}{\partial t}\mathrm{v_t}\left(t, \hat{X}\right) = \frac{\partial}{\partial t}\hat{X}^t \tag{1476}$$

$$\frac{\partial}{\partial t}\hat{X}^t = \hat{X}^t \log\left(\hat{X}\right) \tag{1477}$$

$$\hat{X} + \frac{\partial}{\partial t}\hat{X}^t = \hat{X} + \hat{X}^t \log\left(\hat{X}\right) \tag{1478}$$

D.4.5 Derivation 17 SR

Correct

$$y(A_x) = \frac{1}{A_x} \tag{1479}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1480}$$

$$\int y(A_x)dA_x = \varepsilon_0 + \log(A_x) \tag{1481}$$

$$\int \frac{1}{A_x}dA_x = \varepsilon_0 + \log(A_x) \tag{1482}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \varepsilon_0 + \log(A_x) - \frac{x}{A_x} \tag{1483}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1484}$$

FLAN-T5

21.2 22.5 40.8 22.3
0.79 0.78 0.53 0.78

$$y(A_x) = \frac{1}{A_x} \tag{1485}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1486}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \int \frac{1}{A_x}dA_x - \frac{x}{A_x} \tag{1487}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) \tag{1488}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1489}$$

GPT-4

-10.6 -30.6 -35.7 -24.3
0.76 0.64 0.97 0.64

$$y(A_x) = \frac{1}{A_x} \tag{1490}$$

$$\int y(A_x)dA_x = \int \frac{1}{A_x}dA_x \tag{1491}$$

$$\int \frac{1}{A_x}dA_x = \log(A_x) + \varepsilon_0 \tag{1492}$$

$$\frac{x}{A_x} = \frac{x}{A_x} \tag{1493}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \log(A_x) + \varepsilon_0 - \frac{x}{A_x} \tag{1494}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1495}$$

ChatGPT

-6.2 -12.5 -21.0 -9.5
0.83 0.80 0.82 0.80

$$y(A_x) = \frac{1}{A_x} \tag{1496}$$

$$\int \frac{1}{A_x}dA_x = \log(A_x) + \varepsilon_0 \tag{1497}$$

$$\int \frac{1}{A_x}dA_x - \frac{x}{A_x} = \log(A_x) + \varepsilon_0 - \frac{x}{A_x} \tag{1498}$$

$$\frac{\partial}{\partial x}(\int \frac{1}{A_x}dA_x - \frac{x}{A_x}) = \frac{\partial}{\partial x}(\varepsilon_0 + \log(A_x) - \frac{x}{A_x}) \tag{1499}$$

D.4.6 Derivation 18 SR

Correct

$$u(\lambda) = \sin(\lambda) \tag{1500}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{1501}$$

$$\int u(\lambda) d\lambda = n - \cos(\lambda) \tag{1502}$$

$$\int \sin(\lambda) d\lambda = n - \cos(\lambda) \tag{1503}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{1504}$$

FLAN-T5

8.2 14.1 2.1 14.0
0.88 0.81 0.90 0.81

$$u(\lambda) = \sin(\lambda) \tag{1505}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{1506}$$

$$-\frac{\int u(\lambda) d\lambda}{\cos(\lambda)} = -\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} \tag{1507}$$

$$-\frac{\int u(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{1508}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{1509}$$

GPT-4

13.1 21.3 24.7 17.0
0.87 0.78 0.81 0.79

$$u(\lambda) = \sin(\lambda) \tag{1510}$$

$$\int u(\lambda) d\lambda = \int \sin(\lambda) d\lambda \tag{1511}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{1512}$$

ChatGPT

18.0 -0.5 24.6 -0.7
0.78 0.82 0.64 0.82

$$u(\lambda) = \sin(\lambda) \tag{1513}$$

$$\int \sin(\lambda) d\lambda = -\cos(\lambda) + C \tag{1514}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{-\cos(\lambda) + C}{\cos(\lambda)} \tag{1515}$$

$$-\frac{\int \sin(\lambda) d\lambda}{\cos(\lambda)} = -\frac{n - \cos(\lambda)}{\cos(\lambda)} \tag{1516}$$

$$where\ n = C + 1. \tag{1517}$$

D.4.7 Derivation 19 SR

Correct

$$J(\phi_1) = \sin(\phi_1) \tag{1518}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{1519}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{1520}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1521}$$

$$\sin(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1522}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{1523}$$

FLAN-T5

13.6 26.0 28.6 20.2
0.85 0.72 0.90 0.75

$$(\phi_1) = \sin(\phi_1) \tag{1524}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{1525}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \frac{d}{d\phi_1} \sin(\phi_1) \tag{1526}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{1527}$$

GPT-4

2.2 4.3 -2.5 3.1
0.91 0.89 0.93 0.89

$$J(\phi_1) = \sin(\phi_1) \tag{1528}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \frac{d}{d\phi_1} \sin(\phi_1) \tag{1529}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{1530}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{1531}$$

ChatGPT

10.4 24.6 30.5 17.5
0.85 0.71 0.92 0.74

$$J(\phi_1) = \sin(\phi_1) \tag{1532}$$

$$\frac{d}{d\phi_1} J(\phi_1) = \cos(\phi_1) \tag{1533}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = \sin(\phi_1) \cos(\phi_1) \tag{1534}$$

$$J(\phi_1) \frac{d}{d\phi_1} J(\phi_1) = J(\phi_1) \cos(\phi_1) \tag{1535}$$

D.4.8 Derivation 20 SR

Correct

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1536}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1537}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1538}$$

$$\cos\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1539}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{1540}$$

FLAN-T5

-5.9 -10.7 -9.1 -8.0
0.82 0.83 0.94 0.83

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1541}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1542}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{1543}$$

GPT-4

3.8 -3.9 -3.2 -2.6
0.90 0.90 0.93 0.90

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1544}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \tag{1545}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) \tag{1546}$$

$$\sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{1547}$$

ChatGPT

-8.8 -10.0 -1.7 -7.2
0.83 0.83 0.97 0.83

$$\mathrm{f}'\left(\mathbf{J}_f\right) = \frac{d}{d\mathbf{J}_f} \sin\left(\mathbf{J}_f\right) \tag{1548}$$

$$\mathrm{f}'\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \cos\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) \tag{1549}$$

$$\cos\left(\mathbf{J}_f\right) \sin\left(\mathbf{J}_f\right) = \sin\left(\mathbf{J}_f\right) \cos\left(\mathbf{J}_f\right) \tag{1550}$$

D.4.9 Derivation 21 SR

Correct

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1551}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1552}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1553}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + e^{\Psi_\lambda} \tag{1554}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1555}$$

FLAN-T5

12.4 24.7 8.3 22.2
0.88 0.75 0.91 0.78

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1556}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1557}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1558}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1559}$$

GPT-4

5.9 23.3 16.0 18.3
0.83 0.71 0.85 0.74

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1560}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1561}$$

$$\int e^{\Psi_\lambda} d\Psi_\lambda = \chi + e^{\Psi_\lambda} \tag{1562}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1563}$$

ChatGPT

-0.5 -10.1 -2.8 -10.2
0.87 0.78 1.08 0.80

$$\mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1564}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = e^{\Psi_\lambda} \tag{1565}$$

$$\frac{d}{d\Psi_\lambda} \mathbb{I}(\Psi_\lambda) = \mathbb{I}(\Psi_\lambda) \tag{1566}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \int e^{\Psi_\lambda} d\Psi_\lambda \tag{1567}$$

$$\int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = e^{\Psi_\lambda} + \chi \tag{1568}$$

$$\Psi_\lambda + \int \mathbb{I}(\Psi_\lambda) d\Psi_\lambda = \Psi_\lambda + \chi + \mathbb{I}(\Psi_\lambda) \tag{1569}$$

D.4.10 Derivation 22 SR

Correct

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1570}$$

$$\int \mathrm{y}'(\mathbf{s}) d\mathbf{s} = \int \log(\mathbf{s}) d\mathbf{s} \tag{1571}$$

$$\int \mathrm{y}'(\mathbf{s}) d\mathbf{s} = \mathbf{s}\log(\mathbf{s}) - \mathbf{s} + \omega \tag{1572}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s}) d\mathbf{s} \tag{1573}$$

$$a(\mathbf{s}) = -\mathbf{s}\log(\mathbf{s}) + \mathbf{s} - \omega + \mathrm{y}'(\mathbf{s}) \tag{1574}$$

FLAN-T5

-2.5 -13.1 12.4 -13.1
0.87 0.80 0.71 0.81

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1575}$$

$$\int \mathrm{y}'(\mathbf{s}) d\mathbf{s} = \int \log(\mathbf{s}) d\mathbf{s} \tag{1576}$$

$$\mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s}) d\mathbf{s} = \mathrm{y}'(\mathbf{s}) - \int \log(\mathbf{s}) d\mathbf{s} \tag{1577}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s}) d\mathbf{s} \tag{1578}$$

$$a(\mathbf{s}) = -\mathbf{s}\log(\mathbf{s}) + \mathbf{s} - \omega + \mathrm{y}'(\mathbf{s}) \tag{1579}$$

GPT-4

11.4 12.4 11.5 12.3
0.89 0.88 0.87 0.88

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1580}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s}) d\mathbf{s} \tag{1581}$$

$$a(\mathbf{s}) = \log(\mathbf{s}) - \int \log(\mathbf{s}) d\mathbf{s} \tag{1582}$$

$$\int \log(\mathbf{s}) d\mathbf{s} = -\mathbf{s}\log(\mathbf{s}) + \mathbf{s} - \omega \tag{1583}$$

$$a(\mathbf{s}) = -\mathbf{s}\log(\mathbf{s}) + \mathbf{s} - \omega + \mathrm{y}'(\mathbf{s}) \tag{1584}$$

ChatGPT

-0.2 -7.5 4.6 -7.6
0.80 0.80 0.68 0.81

$$\mathrm{y}'(\mathbf{s}) = \log(\mathbf{s}) \tag{1585}$$

$$\frac{d}{d\mathbf{s}} \mathrm{y}'(\mathbf{s}) = \frac{d}{d\mathbf{s}} \log(\mathbf{s}) \tag{1586}$$

$$\mathrm{y}''(\mathbf{s}) = \frac{1}{\mathbf{s}} \tag{1587}$$

$$a(\mathbf{s}) = \mathrm{y}'(\mathbf{s}) - \int \mathrm{y}'(\mathbf{s}) d\mathbf{s} \tag{1588}$$

$$a(\mathbf{s}) = \log(\mathbf{s}) - \int \log(\mathbf{s}) d\mathbf{s} \tag{1589}$$

$$a(\mathbf{s}) = -\mathbf{s}\log(\mathbf{s}) + \mathbf{s} - \omega + \mathrm{y}'(\mathbf{s}) \tag{1590}$$

D.4.11 Derivation 23 SR

Correct

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1591}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{1592}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1593}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1594}$$

$$\mathbf{v}(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1595}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = (\int \sin\left(F_N\right) dF_N)^2 \tag{1596}$$

$$(\int \mathrm{A_z}\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1597}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1598}$$

FLAN-T5

-2.9 -13.9 -1.8 -10.5
0.92 0.84 0.97 0.84

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1599}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{1600}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1601}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1602}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1603}$$

$$\mathbf{v}^2(F_N) = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1604}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1605}$$

GPT-4

15.9 27.4 34.1 18.8
0.81 0.66 0.78 0.70

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1606}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1607}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1608}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1609}$$

ChatGPT

-14.3 -13.1 -9.9 -10.6
0.78 0.78 0.76 0.78

$$\mathrm{A_z}\left(F_N\right) = \sin\left(F_N\right) \tag{1610}$$

$$\mathbf{v}(F_N) = (\int \mathrm{A_z}\left(F_N\right) dF_N)^2 \tag{1611}$$

$$\int \mathrm{A_z}\left(F_N\right) dF_N = \int \sin\left(F_N\right) dF_N \tag{1612}$$

$$\mathbf{v}(F_N) = (\int \sin\left(F_N\right) dF_N)^2 \tag{1613}$$

$$\mathbf{v}(F_N) = Q^2 - 2Q\cos\left(F_N\right) + \cos\left(F_N\right)^2 \tag{1614}$$

$$Q = \int \sin\left(F_N\right) dF_N \tag{1615}$$

$$(\int \sin\left(F_N\right) dF_N)^2 = \left(Q - \cos\left(F_N\right)\right)^2 \tag{1616}$$

D.4.12 Derivation 24 SR

Correct

$$\mathrm{f}'(\varepsilon_0) = \sin(\varepsilon_0) \tag{1617}$$

$$0 = -\mathrm{f}'(\varepsilon_0) + \sin(\varepsilon_0) \tag{1618}$$

$$\frac{d}{d\varepsilon_0}0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'(\varepsilon_0) + \sin(\varepsilon_0)) \tag{1619}$$

$$0 = \cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) \tag{1620}$$

$$\int 0 d\varepsilon_0 = \int (\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0)) d\varepsilon_0 \tag{1621}$$

FLAN-T5

0.0 0.0 0.0 0.0
1.00 1.00 1.01 1.00

$$\mathrm{f}'(\varepsilon_0) = \sin(\varepsilon_0) \tag{1622}$$

$$0 = -\mathrm{f}'(\varepsilon_0) + \sin(\varepsilon_0) \tag{1623}$$

$$\frac{d}{d\varepsilon_0}0 = \frac{d}{d\varepsilon_0}(-\mathrm{f}'(\varepsilon_0) + \sin(\varepsilon_0)) \tag{1624}$$

$$0 = \cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) \tag{1625}$$

$$\int 0 d\varepsilon_0 = \int (\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0)) d\varepsilon_0 \tag{1626}$$

GPT-4

0.0 1.5 3.9 1.6
0.97 0.98 0.80 0.98

$$\mathrm{f}'(\varepsilon_0) = \sin(\varepsilon_0) \tag{1627}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) = \cos(\varepsilon_0) \tag{1628}$$

$$\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) = 0 \tag{1629}$$

$$\int 0 d\varepsilon_0 = \int (\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0)) d\varepsilon_0 \tag{1630}$$

ChatGPT

18.8 23.4 22.8 23.4
0.77 0.68 0.70 0.68

$$\mathrm{f}'(\varepsilon_0) = \sin(\varepsilon_0) \tag{1631}$$

$$\int \mathrm{f}'(\varepsilon_0) d\varepsilon_0 = -\cos(\varepsilon_0) + C \tag{1632}$$

$$\frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) = \cos(\varepsilon_0) \tag{1633}$$

$$\int \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0) d\varepsilon_0 = \mathrm{f}'(\varepsilon_0) + C \tag{1634}$$

$$\int (\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0)) d\varepsilon_0 = -\sin(\varepsilon_0) + \mathrm{f}'(\varepsilon_0) + C \tag{1635}$$

$$\int 0 d\varepsilon_0 = \int (\cos(\varepsilon_0) - \frac{d}{d\varepsilon_0}\mathrm{f}'(\varepsilon_0)) d\varepsilon_0 \tag{1636}$$

D.4.13 Derivation 25 SR

Correct

$$y(W, q, B) = W + \frac{q}{B} \tag{1637}$$

$$0 = W - y(W, q, B) + \frac{q}{B} \tag{1638}$$

$$\frac{d}{dq}0 = \frac{\partial}{\partial q}(W - y(W, q, B) + \frac{q}{B}) \tag{1639}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B} \tag{1640}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{1641}$$

FLAN-T5

15.9 15.5 18.8 15.2
0.79 0.81 0.71 0.81

$$y(W, q, B) = W + \frac{q}{B} \tag{1642}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{\partial}{\partial q}(W + \frac{q}{B}) \tag{1643}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{\partial}{\partial q}(W + \frac{q}{B}) \tag{1644}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{1645}$$

GPT-4

4.9 10.5 16.3 9.5
0.83 0.75 0.84 0.77

$$y(W, q, B) = W + \frac{q}{B} \tag{1646}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{\partial}{\partial q}(W + \frac{q}{B}) \tag{1647}$$

$$\frac{\partial}{\partial q}(W + \frac{q}{B}) = \frac{1}{B} \tag{1648}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B} \tag{1649}$$

ChatGPT

9.8 20.3 -3.6 20.5
0.77 0.64 0.91 0.64

$$\frac{\partial}{\partial B}(W + \frac{q}{B})y(W, q, B) = W + \frac{q}{B} \tag{1650}$$

$$\frac{\partial}{\partial q}y(W, q, B) = \frac{1}{B} \tag{1651}$$

$$\frac{\partial}{\partial B}y(W, q, B) = -\frac{q}{B^2} + \frac{1}{B} \tag{1652}$$

$$0 = -\frac{\partial}{\partial q}(W + \frac{q}{B}) + \frac{1}{B}\frac{\partial}{\partial B}(W + \frac{q}{B}) \tag{1653}$$

$$0 = -\frac{\partial}{\partial q}y(W, q, B) + \frac{1}{B}\frac{\partial}{\partial B}y(W, q, B) \tag{1654}$$

E Generate Annotated Derivations

Algorithm 1 relies on a premise generation algorithm defined elsewhere (Meadows et al., 2023) alongside other procedures and hyperparameters. The particular hyperparameters used in this work are p_history=10, p_arity_0=5, p_renaming=1, p_arity_1=50, p_evaluate=50, p_arity_2=100, p_int_or_diff=1, p_subs=5. The key difference is that EXTRACT_DERIVATION has been improved, and is called as soon as a new equation is derived, rather than after the derivation reaches a specific length. This (and separately, a better implementation achieved after making this change) improves runtime when creating derivations of a specific length. LENGTH in this case is a sampled integer from a truncated Gaussian described in the Data Analysis section.

Algorithm 1 Derivation Generation Algorithm

procedure DERIVATION($prior\_derivation$)
  if $prior\_derivation$ is None **then**
    $eq \leftarrow$ GET_PREMISE($symbols$)
    $D \leftarrow [(eq, \text{"}premise\text{"})]$
  else
    $D \leftarrow prior\_derivation$
  end if
  $L \leftarrow$ LENGTH
  $count \leftarrow 0$
  while True **do**
    $next\_step \leftarrow$ STEP(D)
    if $next\_step$ is False **then**
      $count \leftarrow count + 1$
    end if
    if $count$ is 100 **then**
      return None
    end if
    $eval\_ints \leftarrow$ [i[1] for i in D if 'evaluate_integrals' in str(i[1])]
    if $next\_step$ is not False and $next\_step$ not in D and $next\_step[1]$ not in $eval\_ints$ **then**
      $D.append(next\_step)$
      $actual\_derivation \leftarrow$ EXTRACT_DERIVATION(D)
      if $length(actual\_derivation) \geq L$ **then**
        break
      end if
    end if
  end while
  return $actual\_derivation$
end procedure
procedure STEP(D; hyperparameters)
  $A \leftarrow$ [i[1] for i in D] ▷ only annotations
  $D \leftarrow$ [i[0] for i in D] ▷ only equations
  Initialize $rules\_0$, $rules\_1$, $rules\_2$.
  Extract relevant equation elements from D into $relevant\_equation\_elements$ for use in $rules$
  Select arity randomly.
  if $arity$ is 0 **then**
    Choose rule from $rules\_0$ and apply it based on conditions.
  else if $arity$ is 1 **then**
    Choose rule from $rules\_1$ and apply it based on conditions.
  else if $arity$ is 2 **then**
    Choose rule from $rules\_2$ and apply it based on conditions.
  end if
  if eq is $sp.Equality$ and it meets certain conditions **then**
    return $(eq, annotation)$
  else
    return False
  end if
end procedure